INVESTIGATING VALIDITY ACROSS TWO TEST FORMS OF THE EXAMINATION FOR THE CERTIFICATE OF PROFICIENCY IN ENGLISH (ECPE): A MULTI-GROUP STRUCTURAL EQUATION MODELING APPROACH

by

Yoko Saito Ameriks

Dissertation Committee:

Professor James E. Purpura, Sponsor

Professor William Snyder

Approved by the Committee on the Degree of Doctor of Education

Date FEB 0 9 2009

Submitted in partial fulfillment of the requirements
for the Degree of Doctor of Education in
Teachers College, Columbia University

2009

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

INVESTIGATING VALIDITY ACROSS TWO TEST FORMS OF THE EXAMINATION FOR THE CERTIFICATE OF PROFICIENCY IN ENGLISH (ECPE): A MULTI-GROUP STRUCTURAL EQUATION MODELING APPROACH

Yoko Saito Ameriks

The purpose of this study was to investigate the comparability of the underlying trait structure of the two test forms of the grammar/cloze/vocabulary/reading (GCVR) section of the Examination for the Certificate of Proficiency in English (ECPE). There were two parts to the study: The first part of the study separately investigated the hypothesized underlying constructs of lexico-grammatical knowledge and reading ability as measured by the GCVR section in the two different forms, using confirmatory factor analysis. The second part examined the extent to which the underlying constructs were invariant across the two different test forms when modeled simultaneously. A multi-group structural equation modeling technique was employed for this part of the study.

The results of the first part of the study showed that both test forms produced identical factorial models. In other words, the same model fit the data well for each form. Moreover, the parameter estimates were very similar in both forms, which suggested that the two forms were comparable when modeled separately.

When the underlying trait structure of each model was estimated simultaneously, the results showed that the invariance across forms was not supported in the data. In other

words, when the equality constraints were imposed on the parameters, the differences in the parameters were found to be statistically significant.

Although the differences in the parameter estimates were statistically significant, the differences in many of the parameters appear to be marginal and the parameter estimates were substantively equivalent. Considering the large sample size of the data used in the study, even a tiny numerical difference in the parameter estimates can result in statistically significant difference. Based on the results of a multi-group SEM, it cannot be concluded that these two forms are strictly equivalent when they are simultaneously modeled. However, it may be reasonable to suggest that the two forms have the identical underlying trait structures despite the marginal differences in the parameter estimates.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

1.1 Context of the Problem

Test scores are often used as a piece of information in making decisions on three different levels: (1) the individual level, (2) institutional level, and (3) public policy level (Kolen & Brennan, 2004). At the individual level, the test scores may, for example, help an individual to decide which college to apply for or which classes to take. Other decisions are made at the institutional level, such as a university deciding on the cut-off test score to use as one of the qualifications for admitting students into their institution. For other decisions, the focus is at a public policy level. For instance, a state board of education looks at the average test scores of the state and attempts to make improvements in the state education system. Regardless of the level of the decision that is to be made, it should be based on the most accurate information possible. To avoid making inappropriate decisions, it is crucial the test scores be accurate representations of the test-takers' ability.

In order to make appropriate decisions so that test scores are an accurate representation of the test-takers' ability, tests are often administered on numerous occasions. For instance, college admissions tests are often offered a few times a year. If identical test items appeared on every test, test-takers may memorize the test items or pass the item-level test information to prospective test-takers. In such situations, test security may be compromised, and a test may be more of a measure of exposure to the

items on the test than of the construct that the test is supposed to measure (Kolen & Brennan, 2004). To prevent this from happening, testing entities often develop and administer multiple test forms. These test forms are based on the same set of test specifications (Millman & Greene, 1989), and test developers follow these specifications to ensure that the test forms are as similar as possible in content and statistical characteristics (Kolen & Brennan, 2004). As a result, the test forms presumably have the same psychometric characteristics, and they are measuring the same underlying constructs. Consequently, test-takers can justifiably take any form of the test since their scores should be comparable across the various forms. This appears theoretically plausible; however, test developers seldom provide empirical evidence that the underlying constructs of each test form are in fact identical. Instead of comparing the underlying constructs of the test forms, test developers often compare the test scores across different test forms. If the test scores are comparable, they tend to presume that the underlying constructs of the test forms are also comparable.

The importance of test form comparability began to gather the attention of testing researchers in the early 1980s (Woldbeck, 1998). Subsequently, the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council for Measurement in Education (NCME) (1999) have revised the standards for educational and psychological testing, and devoted an extensive portion of a chapter of the standards document to the test form comparability issue. Moreover, one of the standards explicitly states the importance of providing evidence of test form comparability:

> A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably. In some cases, direct evidence of score equivalence may be provided. In other cases, evidence may come from a demonstration that the theoretical assumptions underlying procedures for establishing score comparability have been sufficiently satisfied (AERA et al., 1999, p. 57).

According to the standards, there are two ways to show that scores on two different test forms are equivalent. One way is to place scores from multiple test forms on a common scale (i.e., test equating). The other way is to investigate the underlying trait structure of the test forms. The following paragraphs further describe these two ways to provide evidence of test form comparability.

To compare the scores obtained on different test forms, a statistical procedure called test equating is often carried out. Kolen and Brennan (2004) describe test equating as a procedure for placing the scores from multiple test forms on a common scale to establish a correspondence among the scores. In other words, the raw scores on multiple test forms are converted to scale scores so that the scores have the same meaning regardless of the test form.

The basic method of test equating is called the "classical method," which involves mean, linear, and equipercentile equating.[1] In mean equating, the means of Form X and Form Y are set equal to one another. In linear equating, the scores on Form X are converted so that Form X has the same mean and standard deviations as Form Y. In equipercentile equating, the score distribution of Form X is set equal to the score distribution of Form Y by scoring the two tests as percentages (Felan, 2002). In other words, the equated score on Form X has the same percentile rank as a corresponding

---

[1] For more information on classical method and item response theory, please refer to Angoff, (1971); Hambleton, Swaminathan, & Rogers, (1991); Kolen & Brennan, (2004); Lee, Kolen, Frisbie, & Ankenmann (2001).

score from Form Y. These classical statistical methods are easy to perform; however, there is a major drawback. The scores based on the classical methods are population dependent and they can vary from one group of examinees to another due to their ability level (Woldbeck, 1998).

To overcome the population dependency limitation, many researchers use item response theory (e.g., Felan, 2002; Gao, 2004; Luo, Seow, & Chin, 2001; Kim & Hanson, 2002) to equate tests. Due to the invariant property of item response theory (IRT), the item parameters are independent of the ability level of test-takers. At the same time, the ability parameter of a test-taker is invariant across test items (Hambleton, Swaminathan, & Rogers, 1991). In other words, the interpretation of item difficulty and test-takers' ability is consistent (Felan, 2002). In order to perform IRT, the tests must somehow be linked (i.e., have common items in the tests) to permit the scaling of item parameters. Because IRT is a much more useful tool than classical method due to its probabilistic characteristic than its deterministic characteristic, IRT test equating has become standard use of IRT.

Regardless of the method used for test equating, ensuring the comparability of test scores is a critical part of the validation process. It is also important to ensure that the underlying constructs of various test forms are comparable (AERA et al., 1999). Otherwise, the decisions made based on the test scores may not be fair and appropriate. Hence, it is important that the test developers not only investigate the comparability of test scores, but also the underlying constructs of the test forms in order to provide evidence of validity.

To examine the underlying constructs of a test, researchers have often used a statistical procedure called structural equation modeling (SEM) (e.g., Beglar, 2000; Chang, 2004; Kunnan, 1995, 1998; Lee, 2005; Purpura, 1997, 1998, 1999; Sasaki, 1993; Schoonen, 2005; Shiotsu & Weir, 2007; Tsai, 2004; Xi, 2005; Yun, 2005). SEM can be used to investigate the underlying structure of a measure in terms of both observed and latent variables. Furthermore, it can be used to examine the relationship among latent variables. When investigating test comparability using SEM, the first step is to examine the underlying constructs of the individual test forms, as represented by a separate model for each of the two tests. Then, the models are compared to determine the degree to which they are similar. In other words, the invariance of the factorial structure of each form is tested by estimating parameters for the two test forms simultaneously. In theory, if the two test forms are measuring the same underlying constructs, the assumption of invariance across both forms should hold.

This type of multi-group SEM has been done in other studies (Bae & Bachman, 1998; Purpura, 1998, 1999; Yun, 2005). For instance, Purpura (1999) investigated the invariance of strategy use across two different groups with the same test format. More specifically, he examined how the model of strategy use and performance differed for the test-takers in a low-ability group as opposed to those in a high-ability group. The results showed that there were both similarities and differences between the low-ability and high-ability groups in how strategy use impacted their performance. This study and others (presented in more detail in the following chapter) have shown that it is possible to compare two different populations using SEM. If two different populations can be

compared using SEM, it seems reasonable and feasible to compare two different test forms with SEM.

The means of examining the comparability of test forms using SEM is different from performing test equating using IRT. However, both procedures can provide evidence that there is no difference in taking either test form X or test form Y. IRT attempts to equate two test forms, by often linking items and producing a common scale, whereas SEM compares the underlying test structure of the two test forms and then simultaneously investigates the invariance of the underlying constructs across the two test forms.

While it is not the objective of the current study to reject the test equating method using IRT, SEM may provide an alternative method to examine tests with multiple forms. In summary, the purpose of the current study is to examine the comparability of two test forms using SEM.

1.2 The Current Study

The Examination for the Certificate of Proficiency in English (ECPE) is an advanced-level ESL examination, which was established in 1953 by the English Language Institute at The University of Michigan. It is administered annually at over 125 test centers in 20 countries (English Language Institute, 2006a). More than 30,000 people take this exam every year. Test-takers who perform competently on this test are awarded a certificate, which is recognized in many countries as evidence of advanced English language skills. The certificate is often used for hiring and promotion of employees in the public and private sectors. The certificate is also used as a part of the requirements for

study abroad program applications or university admissions. Furthermore, the certificate is recognized by some airline companies, travel agents, and international businesses as proof of English language ability (English Language Institute, 2006b).

The ECPE is designed to measure the following language abilities: speaking, listening, writing, reading, and grammar (English Language Institute, 2006a). In order to measure these abilities, the exam consists of four components: speaking, listening, writing, grammar/cloze/vocabulary/reading (GCVR). It is beyond the scope of this study to examine all different tasks of this test; thus, the current study focused on examining the underlying trait structure of the GCVR section.

The current study was not the first to investigate the construct validity of the ECPE (e.g., Johnson, Yamashiro, & Yu, 2003; Saito, 2003; Wagner, 2004). Wagner (2004) investigated the construct validity of the ECPE listening section using a model of second language listening ability based on Buck's (2001) listening construct. It was hypothesized that the items in the extended listening section measure two constructs: (a) items measuring comprehension of explicitly stated spoken information; (b) items measuring comprehension of implicit spoken information. Wagner performed a series of exploratory factor analyses to examine correlation patterns among the items and to investigate the underlying factors of the extended listening section. His results provided only limited empirical evidence in support of the hypothesized two-factor model. Instead, the evidence suggested either a one-factor or a three-factor model would better fit the data. The one-factor model seemed to show that the extended listening section of the ECPE was measuring a single trait (i.e., listening ability), and the three-factor model seemed to illustrate that the items were text dependent.

Johnson et al. (2003) and Saito (2003) examined the construct validity of the grammar/vocabulary/reading (GVR), and multiple-choice (MC) cloze sections in the old format of the ECPE. They were particularly interested in the underlying trait structure of the MC cloze section in relation to the GVR section. Both studies concluded that the cloze task measured the same underlying trait structure as the GVR tasks. As a result of these studies, the test developers of ECPE decided to collapse the MC cloze and the GVR section of the test into one section.

Although these studies have contributed to understanding the underlying trait structure of a single test form, no study has investigated the comparability of the trait structure of ECPE across different test forms. Since the test forms are assumed to be equivalent based on development from a common set of specifications and IRT test equating procedures, it is also assumed that all forms measure the same underlying trait structure.

At the English Language Institute, researchers examine the comparability of the ECPE test forms using the anchor-test design of item response theory (IRT) (English Language Institute, 2006c). An investigation comparing the invariance of the underlying trait structure of two test forms using SEM, along with the test equating procedure, may provide further evidence of validity of the ECPE.

1.3 Purpose of the Study

The main goal of the present study was to investigate the comparability of the underlying trait structure of two test forms of the grammar/cloze/vocabulary/reading (GCVR) section of the Examination for the Certificate of Proficiency in English (ECPE).

There were two parts to the study: The first part focused on the construct validity of the GCVR section in two forms of the test as it was the initial step in validation to provide evidence that the tasks on the test have measured the constructs which they were designed to measure (Messick, 1993). Based on the preliminary analysis on the GCVR section of the ECPE (Saito, 2003), the present study first separately investigated the hypothesized underlying constructs of lexico-grammatical knowledge and reading ability as measured by the GCVR section in the two different forms.

The second part of the study focused on the comparability of the two test forms when modeled simultaneously. In other words, it examined the extent to which the underlying constructs were invariant across the two different test forms. To do so, a multi-group model including both test forms was formulated so that these models could be simultaneously estimated.

1.4 Research Questions

The current study addressed the following research questions:

1. What is the underlying trait structure of lexico-grammatical knowledge and reading ability as measured by the GCVR section in Form X of the ECPE?
2. What is the underlying trait structure of lexico-grammatical knowledge and reading ability as measured by the GCVR section in Form Y of the ECPE?
3. To what extent does the GCVR section measure the same underlying trait structure across the different ECPE test forms?

1.5 Definition of Key Terms

This section defines and explains the theoretical constructs under investigation, as well as the main statistical procedures used in this study.

1.5.1 L2 Lexico-grammatical Knowledge

According to Purpura (2004), L2 lexico-grammatical knowledge is concerned with a mental representation of what L2 learners know about the form and meaning of utterances and written text in their target language. Grammatical form refers to linguistic forms such as correct formation of words, phrases, and sentences. Grammatical meaning, on the other hand, refers to the literal and intended meaning expressed by grammatical forms.

1.5.2 L2 Reading Ability

Reading ability in this study is concerned with the elements of the interactive processing model. According to the interactive model, readers use both lower-level and higher-level processing skills. Lower-level skills include rapid, automatic, and linguistic processing so that the reader can decode unfamiliar words, process syntactic parsing, and examine the part of speech of a particular word (Alderson, 2000; Grabe, 2005; Grabe & Stoller, 2002; Segalowitz, Poulsen, & Komoda, 1991). Higher-level skills include comprehension and interpretation so that the reader can anticipate what happens next in the text and draw on past experiences (Grabe & Stoller, 2002; Segalowitz et al., 1991). Both skills interact either simultaneously or alternately in reading. By using both levels of skills, readers can grasp the main ideas of the passage, understand the details of the text,

make inferences, and speculate on the meaning of unknown words based on the context (Anderson, 1999).

1.5.3 Structural Equation Modeling

Structural equation modeling is a multivariate analytic procedure for representing interrelationships (1) between observed and latent variables, and (2) among latent variables based on substantive theory or previous empirical research (Kim & Mueller, 1978). Each relationship in the model is determined by a set of mathematical equations. Then, the entire model is tested for overall model data fit.

In this study, a statistical procedure outlined by Jöreskog (1993) is followed. A model-generating procedure is used to specify and test an initially hypothesized model of the relationships among observed and latent variables. When the model does not fit the data, it is modified and retested until a model with a statistical fit and a meaningful explanation is found.

1.6 Importance of the Study

This investigation into the comparability of the underlying trait structures across different test forms has a potential of making a number of contributions to the theory, research methodology, and development of test items, as well as to the field of second language testing in general.

From a theoretical perspective, this study provides information on the underlying trait structures of the grammar/cloze/vocabulary/reading (GCVR) section of the Examination for the Certificate of Proficiency in English (ECPE) developed by

University of Michigan. More importantly, this study attempts to determine whether the underlying trait structures of two test forms are invariant. Furthermore, this study uses Purpura's (1997; 2004) theoretical model of grammatical knowledge, which is in line with the purpose of the GCVR section, in order to determine whether his model can be empirically supported. By attempting to model the underlying trait structures of the GCVR section, it not only provides empirical evidence on what aspects of language ability are measured by this test, but also it presents the degree to which the resulting model fits the observed data.

From the perspective of language testing research methodology, this study is meaningful because it utilizes structural equation modeling (SEM) to compare the underlying trait structures of the two different forms of the GCVR section of ECPE. Although SEM has become a common statistical procedure in current studies (e.g., Chang, 2004; Kunnan, 1995, 1998; Park, 2007; Purpura, 1999; Sasaki, 1993), comparability of the underlying structures across different test forms using SEM appears to be an unfamiliar analytical procedure in language assessment.

When examining the comparability of parallel test forms, it is common to utilize the traditional test theory (i.e., examine the differences in the mean, standard deviation, and reliability coefficient) to insure fairness and consistency in the test (Luo et al., 2001). However, more advanced statistical procedures exist. Indeed the most commonly used procedure in test equating for recent studies (e.g., Kim & Hanson, 2002; Luo et al., 2001) has been IRT, which examines the comparability of multiple test forms by linking the tests.[2] Although IRT is an effective tool for test equating, it does not focus on the

[2] IRT links multiple test forms by including some identical items in the tests. IRT allows for systematic missing blocks of data to be analyzed with the assumption that there is enough linkage between and among

underlying trait structures of the test tasks. Rather, the emphasis is on the comparability of the scores.

The use of SEM, on the other hand, examines the extent to which the trait structure models generated for different test forms are comparable. Because there does not appear to be any prior research utilizing SEM to measure comparability of test forms in the language assessment literature, the present study provides a new focus on complementing the IRT test equating procedure.

Finally, from the perspective of test design, this study provides insights into the underlying trait structures of the GCVR of the ECPE. If the study results show that the two test forms are measuring the same underlying constructs, the test developers can provide further evidence of validity to ensure that two test forms are interchangeable. If, however, the study results show that the two forms are different in their underlying trait structures, there may be an issue with test fairness. Variant underlying structures indicate that the two test forms are measuring different constructs. This means that it would be difficult to make appropriate decisions based on the test scores, which can cause problems with fairness in test use (AERA et al., 1999). If the results show that the two forms are different in terms of their underlying structures, the test developers will need to revise the test accordingly.

1.7 Limitations of the Study

The scope of the present study is limited in several respects that could affect the generalizability of the results. With regards to the theoretical model used in this study, there is a possibility that the proposed models in the current study are not appropriate

models to employ for the purpose of this study, although they have been successfully applied in a pilot study (Saito, 2003). The theoretical models of lexico-grammatical knowledge and reading ability proposed in this study are neither exhaustive nor comprehensive. In other words, other approaches to modeling the GCVR section could fit or explain the data better than the ones employed here. The current study does not attempt to incorporate an exhaustive investigation of all the possible statistical approaches to analyzing language test data. This study can only test and potentially reject or fail to reject models posited in this study, given the variables that are measured.

Another limitation with the modeling could be referred to as the *"naming fallacy"* (Kline, 1998, p. 191). Just because a set of factors in the modeling process is assigned a particular name, those names are not necessarily a correct representation of the factors presented in the model. Therefore, the interpretation of the factors presented in the study – specifically the labels placed on those factors – may not be as clear as it appears.

With respect to the generalizability of results, although this study used the data drawn from a highly heterogeneous group of test-takers, it might differ from the population taking a test other than ECPE, thus limiting the findings in their case. The results of the current study should not be generalized beyond the population the ECPE test-takers represent.

Finally, a limitation related to the generalizability of results relates to the nature of test tasks. All the tasks used in this study were selected-response tasks that ask the test-takers to choose the answer. There is no assumption that the results of this study can be applicable to task and item types other than the ones used in the study. In other words,

findings cannot be assumed to be applicable to other types of grammar, vocabulary, and MC cloze test tasks.

1.8 Summary

In this chapter, the need for more research on the comparability of the underlying trait structures across two different test forms was presented, focusing on one particular large-scale test, the Examination for the Certificate of Proficiency in English (ECPE) administered by the University of Michigan. The purpose, content, and significance of the present study, and the research questions of the study were also presented. The limitations of this study were discussed. In the next chapter, the literature related to this study is reviewed.

Chapter II

REVIEW OF THE LITERATURE

The objective of this chapter is to review several strands of literature relevant to the present study. This chapter begins by defining the theoretical construct of the GCVR section of the ECPE. To do so, the literature review focuses on how lexico-grammatical knowledge and reading ability have been defined and conceptualized in the field of language teaching and assessment. This section closes with a review of the purpose of the GCVR section. This review and analysis of both the substantive theories and purpose of the GCVR section of ECPE reveal the hypothesized underlying trait structures of GCVR section.

The second section of the chapter is divided in two parts. The first part focuses on reviewing the structural equation modeling (SEM) studies using a single-group of participants. The next part focuses on the SEM studies using multiple groups of participants. The multi-group studies provide examples of how SEM can be a useful tool in comparing the underlying trait structures of two test forms.

2.1 Theoretical Construct of Lexico-grammatical Knowledge

This section focuses on defining the theoretical construct of the GCVR section of the ECPE. Therefore, the literature review addresses how lexico-grammatical knowledge has been defined in the field of language teaching and assessment. Following the review, theoretical models used for this study are presented.

2.1.1 Componential Approach: Lado and Carroll

A number of researchers have attempted to conceptualize and define grammatical knowledge in models of communicative competence. The first model was presented by Lado (1961) who applied structuralism with behaviourism in psychology and in language teaching to define L2 proficiency. He proposed an "elements and skills" model, which consisted of four representative language skills (i.e., listening, reading, speaking, and writing) and linguistic elements (i.e., pronunciation, grammatical structure, and lexicon). The model asserted that demonstrating knowledge of these linguistic elements in the context of the language skills was sufficient to provide evidence of L2 learners' language proficiency.

In the same year, also influenced by structuralism, Carroll (1961) presented a multi-dimensional-four-skill model, which was similar to what Lado (1961) had proposed. Carroll's model viewed language ability as a composition of skills (i.e., listening, reading, speaking, and writing) and language components (i.e., phonology, morphology, syntax, and lexicon). As shown in Table 2.1, each cell presented an independent language ability and "different kinds of mastery are displayed against different aspects of the language structure" (Carroll, 1961, p. 34). According to Carroll, it was theoretically possible to measure one component of language ability at a time (i.e., a discrete-point approach).

Table 2.1
Multi-Dimensional-Four-Skill Model for Second Language Proficiency
(Adapted from Sasaki, 1999)

		Language Components			
		Phonology / Orthography	Morphology	Syntax	Lexicon
Skills	Listening				
	Reading				
	Speaking				
	Writing				

Carroll (1968) further refined the model of L2 proficiency by dividing the skills into two categories: receptive skills (i.e., listening and reading) and productive skills (i.e., speaking and writing). With this model, he asserted that tests should be designed to predict the use of language elements and skills in future social situations that the test-takers might experience in their life. In other words, L2 proficiency is the ability to demonstrate control of phonology/orthography, morphology, syntax, and lexicon with one of the language skills used in the target language use contexts.

The idea of focusing on the four skills is considered useful, even to this day, by teachers, textbook writers, and test developers. For instance, many language tests distinguish sections in terms of listening, reading, speaking, and writing skills – now referred to as a language use. Also, various language textbooks organize sections in terms of skills, such as, "how to improve listening, reading, speaking, and writing". Even teachers and learners often find it helpful to distinguish between receptive and productive skills and between spoken and written channels (Allison, 1999). Thus, it is clear that the four skills model (Carroll, 1961; Lado, 1961) remains pedagogically useful.

Despite the practical use of the model, it had serious shortcomings as a principled account of how language ability is defined. What the model failed to consider is some of

the shared characteristics across skills. For example, the lexical choice of the test-taker may be common to both speaking and writing (productive) skills. Moreover, the test-taker's ability to separate important information from trivial information is common to both reading and listening (receptive) skills. Thus, many researchers considered it to be problematic to measure one aspect of language at a time. Instead, they began to focus more on "integrative" approach, that is, to measure several modalities at the same time. An example of an integrative approach is a test which allows test-takers to use reading, listening, phonology, morphology, syntax, and lexicon in an active interplay with spoken or written discourse.

2.1.2 Holistic Approach: Spolsky

In the early 1970s, the way to define L2 proficiency shifted focus from a discrete-point approach to an integrative approach. Spolsky (1973) focused on measuring test-takers' overall proficiency rather than the specific language components. He stated, "We must try to find some way to get beyond the limitation of testing a sample of surface features, and seek rather to tap underlying competence" (Spolsky, 1973, p. 175). A similar statement was made by Brière (1971) who also recognized the limitations in the discrete-point approach: "The language tests being used today are limited to measuring what is on the 'surface,' and can give us no information about what is underneath" (p. 385). Owing to these scholars, the notion of the separability of the language skills to define L2 proficiency faded and the idea of a general language proficiency factor began to emerge.

2.1.3 Unitary Competence Hypothesis: Oller

Oller (1979) was an early influential advocate of the unidimensionality of language proficiency. He claimed that the different linguistic components are so closely interrelated that they are essentially unitary (i.e., unitary competence hypothesis). Oller attempted to elaborate the notion of unitary competence hypothesis based on the redundancy of language. As Barnwell (1996) stated, "redundancy permeates languages, and to know a language is some way to know its redundancies" (p. 108). In other words, redundancy helps comprehension when the means of communication are not clear. For instance, if two people are in a crowded room with a loud noise, they can still interact adequately because redundancy provides enough contextual clues to prevent misunderstanding in the communication. On the other hand, a telephone number provides no redundancy. If one digit is misheard, then the entire message becomes meaningless because each digit carries as much information as any other and there is no association among them. Because the ability to use redundancy makes speculations about what is to follow, Oller asserted that the ability to exploit redundancy was one of the characteristics of competence in language.

Based on the notion of redundancy, Oller (1979) proposed the idea of "pragmatic expectancy grammar" which is "a psychologically real system that sequentially orders linguistic elements in time and in relation to extralinguistic elements in meaningful ways" (p. 34). In other words, pragmatic expectancy grammar relates the form of linguistic components to contextual meanings. Based on the idea of redundancy and pragmatic expectancy grammar, Oller argued that a construct of L2 proficiency should consist of a

learner's capability to make surmises about unknown elements by employing the redundancy used in the language.

In order to provide empirical evidence of unitary competence hypothesis, Oller and other researchers conducted a number of studies utilizing principal component analysis (Irvine, Atai, & Oller, 1974; Oller, 1979, 1983a; Oller & Hinofotis, 1980; Scholz, Hendricks, Spurling, Johnson, & Vandenberg, 1980). These studies showed the existence of one major factor to account for most of the common variance in a various language tasks. Based on these results, the researchers were convinced that there was a "general language proficiency factor," known as a G-factor.

Following Oller's notion of a unitary competence hypothesis, many researchers investigated whether language proficiency is "divisible" or "unitary". Contrary to Oller's findings, subsequent research results indicated that language ability was multi-componential (Bachman & Palmer, 1980, 1981a, 1981b; Carroll, 1983; Vollmer & Sang, 1983). Concurrently, researchers began questioning the statistical procedures used in Oller's factor analytic studies (Carroll, 1983; Farhady, 1983; Porter, 1983). For example, Farhady (1983) argued that the principal component analysis Oller utilized failed to isolate the entire variance into error and unique variance. This caused Oller to mistakenly interpret the statistical results as a sign of a unitary factor. By examining all the studies related to this issue, Vollmer and Sang (1983) proposed that there are strong and weak versions of the unitary and divisible trait hypothesis. They further argued that there were no statistically appropriate studies confirming the strong versions of either hypothesis; however, the weak versions had some supporting evidence. At the end of this heated debate, Oller (1983b) admitted that the strong form of the unitary competence hypothesis

was untenable, recognizing that the general factor indicated in his studies may have been inflated due to the use of the principal component analysis.

Although the idea of unitary competence hypothesis was rejected, Oller's contribution to the field was substantial. The importance of discourse in language knowledge and use had been previously introduced by Lado (1961) and Spolsky (1968); however, Oller's concern was not restricted to the aspect of discourse. Instead, his idea of pragmatic expectancy grammar involved cognitive processing, which allowed the learner to make the most efficient use of language knowledge in the given context. In other words, he provided an important insight into the dynamic and interactive nature of second language proficiency (Chang, 2004). Although the statistical procedure Oller used to examine the structure of L2 proficiency turned out to be inappropriate, it led researchers into subsequent studies of construct validation, which has since been a central concern of L2 language testing research (Chang, 2004).

2.1.4 Canale and Swain's Model: Communicative Competence

In the 1980s, L2 proficiency models focused not only on the linguistic components of language, but also on language as communication. The most significant theoretical framework to address issues of language as communication was proposed by Canale and Swain (1980). Before discussing their model of language ability, it is important to comprehend the context within which their model was derived. In the 1960s, Chomsky (1965) asserted that linguistic competence involves language knowledge but not ability for use. Opposing this idea, Hymes (1972) argued that sociolinguistic appropriateness expressed in context is part of language competence. For Hymes,

language is a form of social interaction and language is used as the means of communication. This view was widely accepted in the field of language teaching in the 1970s; however, it was not successfully conceptualized in the language testing field until Canale and Swain (1980) postulated their view of language competence (Chang, 2004).

Canale and Swain (1980) proposed an integrated view of language competence in their model of communicative competence. This model was composed of grammatical competence, sociolinguistic competence, and strategic competence. Grammatical competence included lexicon, morphology, phonology, and sentence-level syntax and meaning. Sociolinguistic competence comprised sociocultural appropriateness rules and discourse rules. Strategic competence encompassed verbal and nonverbal communication strategies used to repair communication breakdowns caused by deficiencies in grammatical and sociolinguistic competence. Later, Canale (1983a, 1983b) viewed sociocultural appropriateness rules and discourse rules as separate competences and created another component of communicative competence, discourse competence. Discourse competence comprised the rules concerning cohesion and coherence of discourse in L2. Table 2.2 is a graphic presentation of Canale's (1983a) model of communicative competence.

Table 2.2
Canale's (1983a) Model of Communicative Competence

Grammatical Competence	Sociolinguistic Competence	Discourse Competence	Strategic Competence
• Lexicon • Morphosyntax • Phonology • Semantics	• Sociocultural appropriateness rules	• Cohesion • Coherence	• Verbal & Non-verbal communication strategies

Canale and Swain (1980) defined grammatical competence in terms of syntax and semantics. Furthermore, they elaborated the interrelated features of grammatical form and grammatical meaning, by including phonology, morphosyntax, lexicon, and semantics in the notion of grammatical competence. However, even though they acknowledged the interrelatedness of form and meaning, they did not articulate the relationship between the two. Correspondingly, they failed to present how the grammatical, sociolinguistic, discourse, and strategic components are associated with one another. Also, while some research has been conducted to validate their model of communicative competence (e.g., Bachman & Palmer, 1982; Harley, Cummins, Swain, & Allen, 1990; Swain, 1985); the studies were unsuccessful in supporting the distinction among grammatical, sociolinguistic, discourse, and strategic competence. For instance, Harley et al. (1990) collected data from 175 test-takers in a French immersion program in Canada. They performed confirmatory factor analysis on the data and found that a single global factor emerged instead of multiple factors. Hence, they could not provide empirical evidence for Canale and Swain's (1980) model of communicative competence.

In spite of these limitations, Canale and Swain's (1980) model of communicative competence made an invaluable contribution to the field of applied linguistics. They provided a comprehensible theoretical model of communicative competence, which significantly promoted communicative trends in language teaching.

2.1.5 Bachman and Palmer's Model

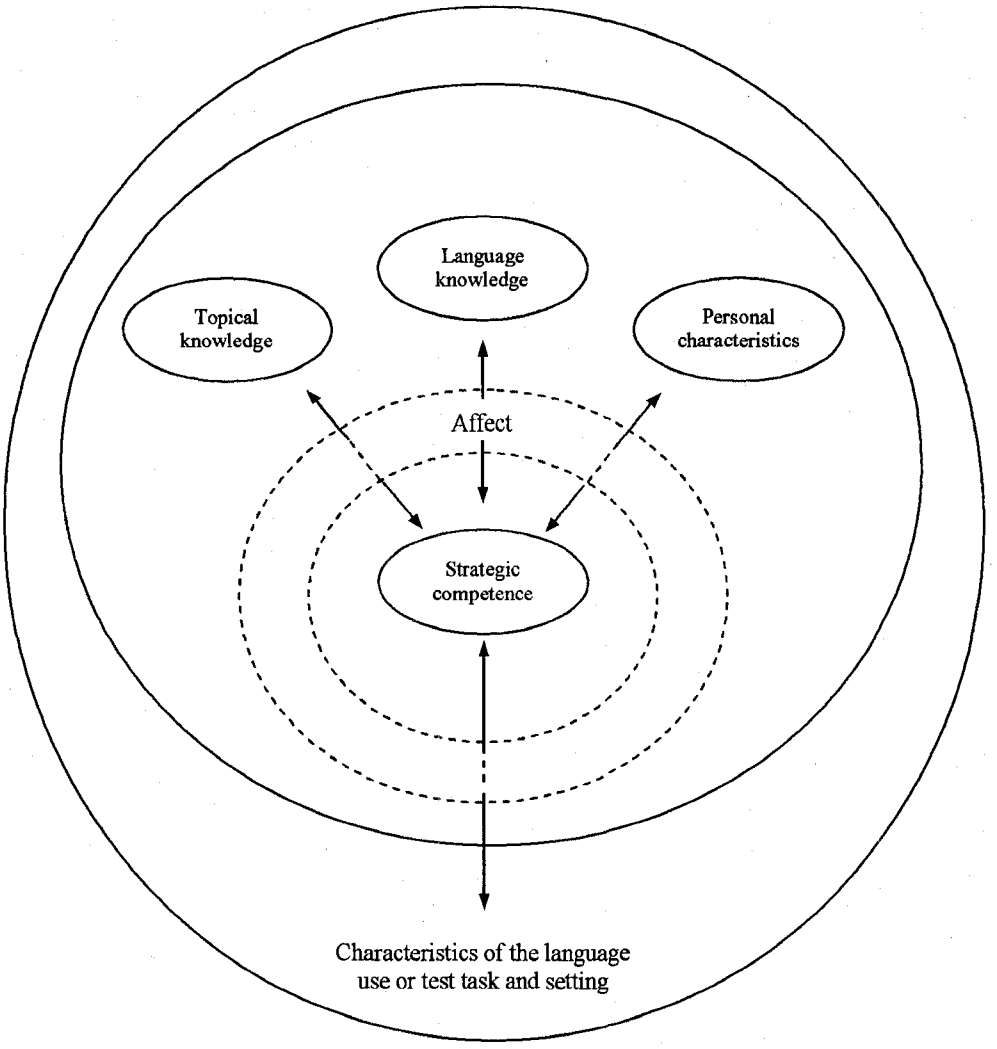
Drawing upon the work of Canale and Swain (1980), Canale (1983a, 1983b), and many others (e.g., Austin, 1962; Halliday 1973, 1976; Halliday, McIntosh, & Strevens,

1964; Hymes, 1972; Searle, 1969; Widdowson, 1978), Bachman (1990) and later Bachman and Palmer (1996) postulated a theoretical framework of communicative language ability, which built upon, but was distinct from that of Canale and Swain's (1980).

First, Bachman and Palmer (1996) characterized strategic competence as a set of metacognitive components, which interact with language knowledge, topical knowledge, personal characteristics, and affect. These components interact with one another so that language users can create and interpret discourse appropriately in a given situation. Canale and Swain (1980) included the notion of strategic competence in their framework; however, they failed to explain how it relates to other components of communicative competence. A visual metaphor of language use and performance on language tests is presented in Figure 2.1.

Second, Bachman and Palmer (1996) included non-linguistic components of communicative language ability in their model. For instance, a test-taker's language knowledge, topical knowledge, and personal characteristics are hypothesized to interact with his/her strategic competence in a given context. In other words, their model perceived language ability as "an internal construct, consisting of language knowledge and strategic competence, that interacts with the language user's topical knowledge and other internal characteristics (e.g., affect), as well as with the characteristics of the context" (Purpura, 2004, p. 54).

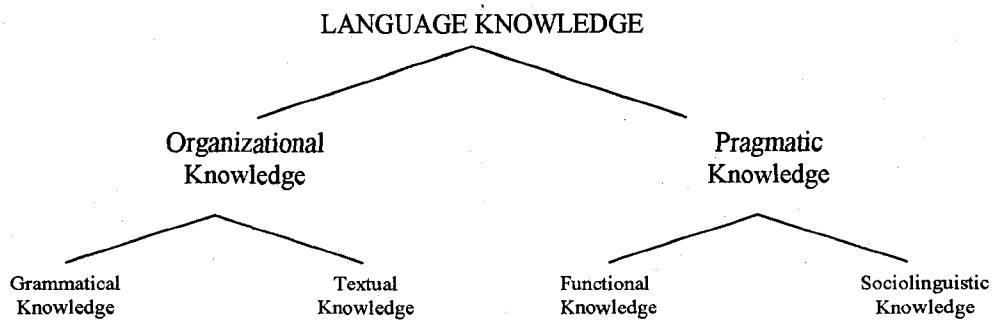
Figure 2.1
Bachman and Palmer's View of Language Use and Language Test Performance
(Adopted from Bachman & Palmer, 1996, p. 63)



According to Bachman and Palmer's (1996) model, communicative competence consists of two components: language knowledge and strategic competence. Language knowledge is further divided into organizational knowledge and pragmatic knowledge (see Figure 2.2). Organizational knowledge is concerned with how individuals control language structure to produce correct formation of words, phrases, and sentences,

whereas pragmatic knowledge deals with how individuals convey meaning and produce sociolinguistically appropriate utterances and sentences (Bachman & Palmer, 1996).

Figure 2.2
Bachman and Palmer's (1996) Model of Language Knowledge



Organizational knowledge is further divided into grammatical knowledge and textual knowledge. Grammatical knowledge is knowledge about phonology, graphology, vocabulary, and syntax, which is required for producing and comprehending accurate utterances and sentences on the sentential level. On the other hand, textual knowledge is concerned with knowledge of cohesion (e.g., conjunction, pronouns, lexical repetition), rhetorical organization (e.g., logical connectors), and conversational organization (e.g., turn-taking). In summary, grammatical knowledge deals with utterances and sentences on the subsentential or sentential level whereas textual knowledge refers to utterances and sentences on the suprasentential and discourse level.

Pragmatic knowledge is divided into functional knowledge and sociolinguistic knowledge. Functional knowledge is concerned with communicative goals of language users and the context in which language is used. In other words, functional knowledge allows individuals to express or interpret intended language functions in a communicative context. For example, when someone asks, "Do you know what time the library opens?",

the utterance usually implicitly includes a request for information on library hours as opposed to a simple 'yes' or 'no' answer. A response such as "Yes, I do", is accurate in terms of the literal meaning of the question; however, it is an inappropriate response as it misinterprets the function of the question as a request for information (Bachman & Palmer, 1996).

Sociolinguistic knowledge allows individuals to utilize situation-specific language (i.e., formal or informal registers) to alter language to a particular language use setting (Purpura, 2004). For example, a professor uses a formal register when delivering a talk at an academic conference while she uses an informal register when talking to her children at home.

The second component in Bachman and Palmer's (1996) theoretical framework is strategic competence, which is referred to as a set of metacognitive strategies individuals use in a communicative language situation. They perceived strategies such as goal setting, assessment, and planning as part of strategic competence, which requires higher order thinking processes. They also proposed that strategic competence interacts with the user's language knowledge, topical knowledge, and personal characteristics in a given context.

This multi-componential model has been widely accepted as a representative of the compensatory and interactive nature of language ability for various reasons. First, this model accounted for non-linguistic components of communicative language ability. It also explained the relationship within and across linguistic and non-linguistic components of communicative language ability.

Second, Bachman and Palmer's (1996) model was based on empirical evidence from their previous (1982) study. Bachman and Palmer (1982) investigated the construct

validity of a test which included an oral interview, a writing test, a multiple-choice grammar test, and a self-rating questionnaire. The test claimed to measure three types of competence: grammatical competence, pragmatic competence, and sociolinguistic competence. They hypothesized that (1) grammatical competence encompassed syntax and morphology, (2) pragmatic competence included vocabulary, cohesion, and organization, (3) sociolinguistic competence consisted of register, nativeness, and non-literal language. They used a multi-trait-multi-method (MTMM) matrix design to analyze the data. They collected data from 116 ESL learners from 18 different language backgrounds and used confirmatory factor analysis to analyze the data. These analyses led them to conclude that sociolinguistic competence was a separate constituent from grammatical and pragmatic competence. Based on this finding, Bachman and Palmer (1996) grouped the components of what they had previously called grammatical competence (i.e., morphology and syntax) and pragmatic competence (i.e., vocabulary, cohesion, and organization) together and renamed it "organizational knowledge" (i.e., how utterances or sentences and texts are organized).

While Bachman and Palmer's (1996) model has been accepted as a comprehensive conceptualization of language ability (Alderson, 1991; McNamara, 1996; Skehan, 1991), some have suggested that further elaboration may be needed (Chang, 2004; McNamara, 1996; Purpura, 1999; 2004). McNamara (1996) and Purpura (1999) argued that the depiction of strategic competence in Bachman and Palmer's (1996) model can be better clarified. Bachman and Palmer acknowledged that strategic competence interacts with language knowledge and affect; however, it remains uncertain as to how affect might be operationalized and how it might relate to different types of strategies

such as metacognitive, cognitive, and social strategies. Furthermore, the depiction of metacognitive strategies in their model is not based on empirical research (Purpura, 1999).

Purpura (2004) argued that from an assessment perspective, Bachman and Palmer's (1996) description of grammatical knowledge defined as form is restricted to sentence level phonology, graphology, vocabulary, and syntax. Bachman and Palmer's depiction of grammatical knowledge is useful if a test developer attempts to measure only linguistic forms. For instance, if a test developer wants to assess a test-takers' knowledge of future tense forms, a discrete-point test of grammar can be created. The test will have questions assessing aspects of the verb form (*will* + present verb form). Although this view of grammatical knowledge defined as form can be useful in some testing situations, it does not account for situations where a test-taker might know the form, but be unclear about the meaning (Purpura, 2004). Furthermore, Bachman and Palmer's definition of grammatical knowledge does not distinguish between the different types of meanings that grammatical forms encode.

Purpura (2004) provides an example to illustrate the situation where this may become an issue:

> Imagine we wanted to determine a student's grammatical knowledge of the simple present, the simple past, and the present perfect tenses as used in conversational narratives. This is a case in which we might wish to test for both grammatical form and meaning, in order to ask questions such as: What makes the three tenses different in terms of time? Does the learner know to use the present perfect to communicate the notion of current relevance in announcing that a story is about to be told? (I've never been more embarrassed!)? Once the story begins, does the learner know to use the past tense to set the scene and the present to tell the sequence of events (We were talking when this waiter appears and uncorks the cava...)? ... All along, the learners could make mistakes that relate to grammatical form and/or grammatical meaning, an analysis of which could inform teachers on how to refocus their teaching and learners on how to direct their learning (pp. 55-56).

Although Bachman and Palmer (1996) included grammatical, textual, functional, and sociolinguistic components in their definition of language knowledge, it is uncertain how these components can be associated with actual language use in a given context (Purpura, 2004). Bachman and Palmer addressed the meaning of language to some extent under organizational knowledge (vocabulary), textual knowledge (cohesion), functional knowledge, and sociolinguistic knowledge; however, given the central role of meaning in language instruction and communicative language use, a more precise illustration of this aspect of language knowledge would be beneficial (Chang, 2004; Purpura, 2004).

In sum, many researchers have attempted to conceptualize communicative language ability by proposing a series of different theoretical models. All the models discussed so far have had a limited focus on how grammatical form might relate to grammatical meaning in communicating literal and intended meanings. In order to fully represent aspects of language ability, the meaning component of grammatical knowledge should also be taken into account.

2.1.6 Rea-Dickins' Model

Rea-Dickins (1991), following Leech (1983)'s idea of communicative grammar, proposed that grammatical knowledge includes not only syntax and semantics, but also pragmatics. She challenged Canale and Swain (1980) and Bachman (1990), arguing that they overlooked the interdependence and interaction among syntax, semantics, and pragmatics. According to Rea-Dickins (1991), communicative grammar requires more than grammatical accuracy. Pragmatics should be taken into account in order to produce semantically acceptable syntactic forms.

Rea-Dickins' (1991) idea of treating pragmatics as part of grammatical knowledge provides a valuable perspective; however, there remain some concerns. First, neither Canale and Swain (1980) nor Bachman (1990) failed to recognize the notion of pragmatics in their models (Purpura, 2004). They treat pragmatics as a separate component from grammatical ability in their frameworks. Second, Rea-Dickins' definition of communicative grammar may have been too broad in that it failed to distinguish between grammar and language (Purpura, 2004). It may be true that syntax, semantics, and pragmatics are interrelated; however, it does not necessarily mean that all three components are constituents of grammatical knowledge.

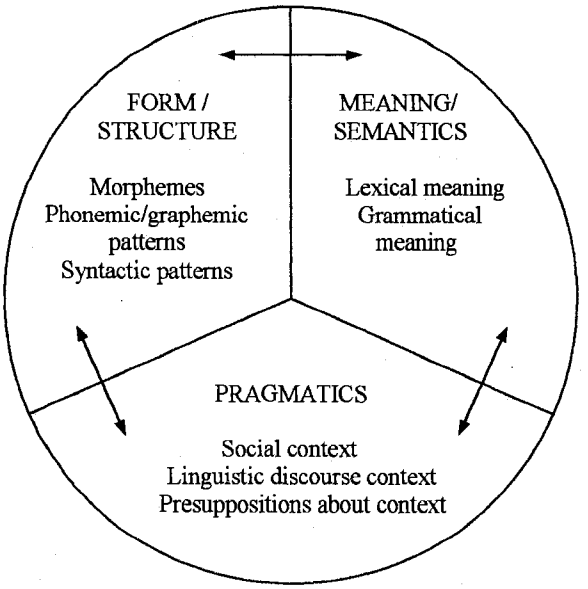
2.1.7 Larsen-Freeman's Model

From a pedagogical perspective, Larsen-Freeman (1991) proposed a similar framework to that of Rea-Dickins' (1991). Larsen-Freeman's model included three interrelated, but separable components: grammatical form, semantic meaning, and pragmatic use. Grammatical form refers to phonological, morphological (i.e., how words are formed), and syntactic forms (i.e., how sentences are formed). Semantic meaning is concerned with the literal meaning encoded in a grammatical structure. The third component, pragmatic use, refers to the lexico-grammatical choices a speaker makes in communication. In other words, pragmatic use explains when and why a certain linguistic pattern is used in a specific context, instead of another pattern with the same literal meaning. Consider the following dialogue:

Situation: Wife telling her husband about a restaurant she went to for lunch.
Wife: I tried the new restaurant on Chestnut Street for lunch, and I really liked it.
Husband: (a) With whom did you go?
(b) Who did you go with?

The husband's utterances (a) and (b) are both grammatically correct and convey the same literal meaning; however, the latter utterance is more appropriate in the context of an informal conversation between a husband and wife. As this example illustrates, it is important to acknowledge the discourse context in which grammatical forms were used to convey meanings. According to Larsen-Freeman (1991), pragmatic use is concerned with social context, linguistic discourse context, and presuppositions about situational context. She emphasized the importance of teaching all three components (i.e., grammatical form, semantic meaning, and pragmatic use) to L2 learners, so that they can utilize linguistic forms accurately, meaningfully, and appropriately. A graphic representation of Larsen-Freeman's framework is presented in Figure 2.3.

Figure 2.3
Larsen-Freeman's Framework for Teaching Grammar
(Adopted from Larsen-Freeman, 1991, p. 280)



Similar to Rea-Dickins' (1991) view, Larsen-Freeman (1991) treated pragmatics as an integral part of grammatical knowledge, which seemed overly generalized for

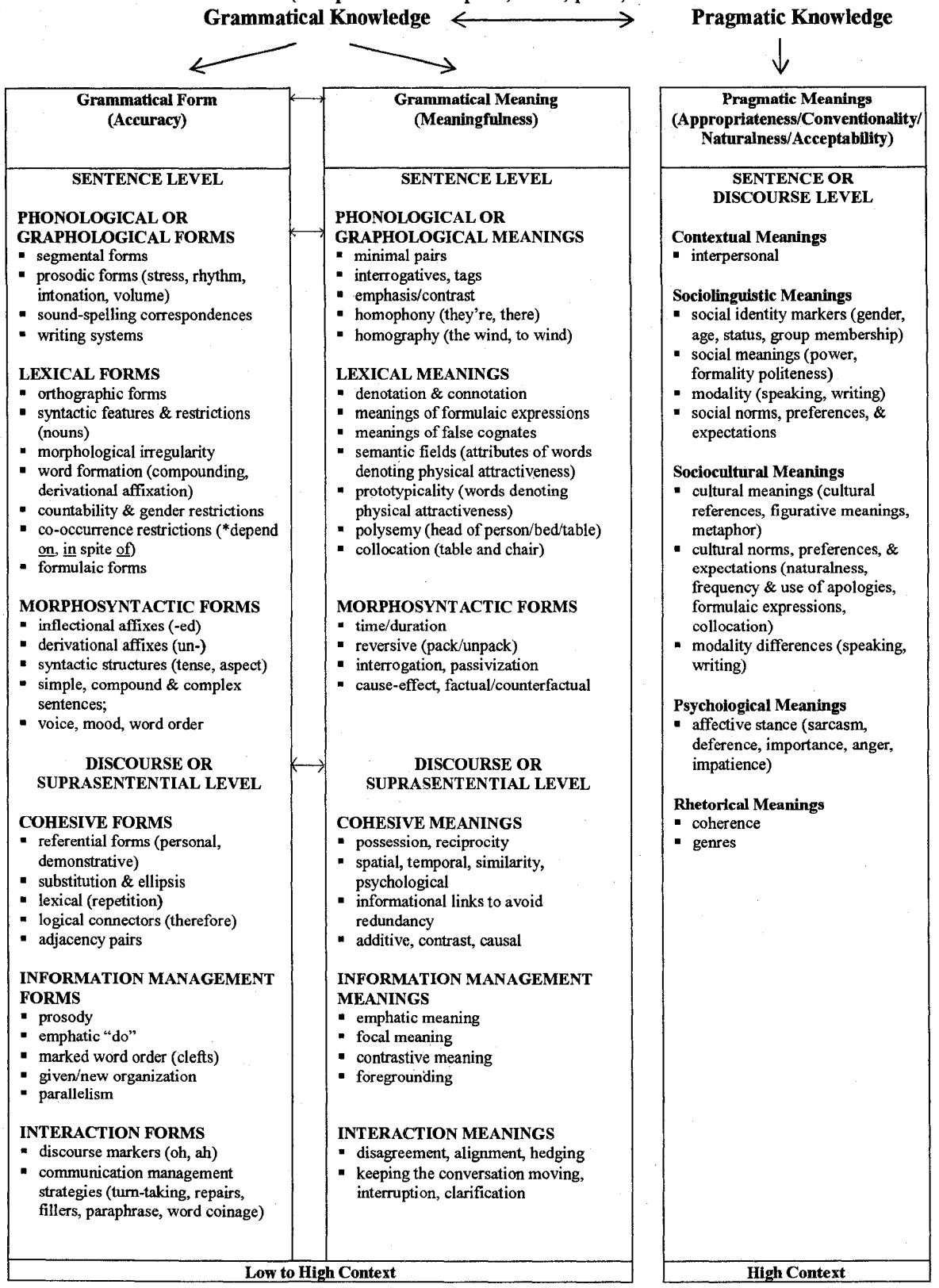
assessment purposes. Both Rea-Dickins and Larsen-Freeman perceived grammar as coterminous with language.

Contrary to their view, Purpura (2004) argued that "there is a fundamental difference in how grammatical forms and meanings are used to evoke literal and intended messages and then how they are used to convey implied meaning that requires pragmatic inference" (p. 60). He further asserted that the inclusion of pragmatic use in the construct of grammar obscured the boundaries between grammar and language. Hence, in Purpura's view, grammatical knowledge and pragmatic knowledge are separate components of language ability. His theoretical model is described in the following section.

2.1.8 Purpura's Model

Building on the work of Canale and Swain (1980), Bachman and Palmer (1996), Rea-Dickins (1991), and Larsen-Freeman (1991), Purpura (2004) posited a theoretical model of grammatical ability for the purpose of instruction and assessment. His model took the complex nature of L2 knowledge into account and treated grammatical and pragmatic knowledge as separate components. In his view, these components are clearly distinct, yet closely related. His model is illustrated in Figure 2.4.

Figure 2.4
Components of Grammatical and Pragmatic Knowledge
(Adopted from Purpura, 2004, p. 91)



The first component of grammatical ability in Purpura's (2004) model is grammatical knowledge. Grammatical knowledge is composed of grammatical form and grammatical or semantic meaning on both the sentence and discourse levels. Grammatical form refers to linguistic forms such as: (1) phonological/graphological forms (e.g., stress: *pro*duce vs. *pro*duce); (2) lexical forms (e.g., countability: *mice* is a plural form of *mouse*); (3) morphosyntactic forms (e.g., affixes: -ed, un-); (4) cohesive form (e.g., personal pronouns: *he*, *she*); (5) information management form (e.g., cleft sentence: *It is Mary who went to New York*); and (6) interactional forms (e.g., discourse markers: *ah*, *oh*). Grammatical meaning, on the other hand, refers to literal and intended meaning expressed by one or more of the grammatical forms listed above. It includes: (1) phonological/graphological meaning (e.g., homophony: *their* vs. *there*); (2) lexical meaning (e.g., collocation: *Merry Christmas* and not **Merry Hanukkah*); (3) morphosyntactic meaning (e.g., reversive: *do* and *undo*); (4) cohesive meaning (e.g., informational links to avoid redundancy: *Where* [should I meet you for lunch]?); (5) information management meaning (e.g., emphatic meaning: *Sam gave a card to Mary* [not Jenny] vs. *Sam gave Mary a card* [not a rose]); and (6) interactional meaning (e.g., repair a conversation: *What do you mean?*).

The second component in Purpura's (2004) theoretical model is pragmatic knowledge, which refers to "a domain of extended meanings which are superimposed upon forms in association with the literal and intended meanings of an utterance" (p. 75). In other words, pragmatics not only refers to literal and intended meaning, but also implied meanings that derive from the context of language use. Pragmatic knowledge encompasses: (1) contextual meanings (e.g., interpersonal meanings); (2) sociolinguistic

meanings (e.g., gender, age, status); (3) sociocultural meanings (e.g., cultural references, figurative meanings); (4) psychological meanings (e.g., sarcasm, criticism, humor); and (5) rhetorical meanings (e.g., coherence, genres).

In summary, Purpura's (2004) framework of language ability consists of two distinct, but related components: grammatical knowledge and pragmatic knowledge. He claimed that if pragmatic knowledge was part of grammatical knowledge, it would be challenging to distinguish the concept of 'grammar' from that of 'language'. Hence, similar to Canale and Swain (1980) and Bachman and Palmer (1996), he treats pragmatic knowledge separately from grammatical knowledge. Purpura's (2004) model corresponds with Bachman and Palmer's (1996) model to a certain degree in that both delineate a distinction between grammar and pragmatics. His model, however, accounts for the difference between form and meaning at both the sentential and suprasentential levels and attempts to differentiate meaning on the semantic and pragmatic levels.

2.1.9 Model for the Current Study: Measuring Lexico-grammatical Knowledge

Many theoretical models defining grammatical knowledge have evolved and been re-conceptualized over time. On the basis of a thorough literature review, Purpura's (2004) model is arguably the most comprehensive model currently available in assessing lexico-grammatical knowledge. Therefore, the present study employed Purpura's model as the model of lexico-grammatical knowledge for the grammar/cloze/vocabulary section of the ECPE.

In order to operationalize the test according to the theoretical model, it is important to specify the purpose of the GCVR section of the ECPE. The test developers

at the English Language Institute of University of Michigan described the purpose of the grammar, MC cloze, and vocabulary tasks as follows:

> Grammar section:
> The grammar items are designed to measure the ability to recognize and select grammatical forms appropriate to convey and interpret explicit and implied meaning appropriate to a specific context. To be specific, this section assesses phonological, graphological, morphosyntactic, and lexical forms, and how these forms combine for semantic and discoursal purposes.
> MC Cloze section:
> The cloze items are developed so that the examinees can demonstrate an ability to read and understand prose texts, and to select, from several options, an appropriate word to fill in a gap.
> Vocabulary section:
> The vocabulary items test advanced level examinees and are designed to measure knowledge of lexis common in academic or business discourse. Some items focus on breadth of lexical knowledge (knowing the meaning of a range of words) and others focus on depth of lexical knowledge (knowing collocations).
> (English Language Institute, 2006c, pp. 7-8)

As highlighted above, the grammar section measures the ability to recognize appropriate grammatical forms to convey explicit and implied meaning in the given context. The vocabulary section measures the ability to recognize the appropriate meaning of words in the given context. These sections together appear to measure form and meaning, which is in line with Purpura's model of grammatical knowledge. Based on a preliminary study (Saito, 2003), the MC cloze section also appears to measure the same trait as the grammar and vocabulary sections. Given the focus of all three sections of the test on grammatical form and meaning, this study can be conceptualized in terms of Purpura's (2004) theoretical model.

2.2 Theoretical Construct of Reading Ability

Just as there have been numerous theories proposed for defining grammatical knowledge, many researchers have attempted to define reading ability. Grabe (1991)

defined reading as a "rapid, purposeful, interactive, comprehending, flexible, and gradually developing" process (p. 378). He argued that fluent reading is *rapid* because the reader needs to sustain sufficient speed while reading to make connection and inferences. Reading is *purposeful* because the reader needs to have a purpose for reading whether it is for leisure, work, or test. Reading is *interactive* because there is an interaction between the reader and the text. Reading is *comprehending* because the reader attempts to understand what the text says. Reading is *flexible* because the reader can use various reading strategies (e.g., skimming, making inferences, skipping function words) to read effectively. Lastly, reading *develops gradually*. The reader does not become a fluent reader overnight. Instead, it takes a long time and requires a significant amount of effort to become fluent in reading.

Grabe's (1991) definition of reading has been adopted by several scholars. According to Anderson (1999), reading is "an interactive, fluent, and active process which involves the reader and the reading material in building meaning" (p. 1). The text does not contain any meaning unless the reader combines the words in text with his/her background knowledge and experience (Bernhardt, 1991; Grabe & Stoller, 2002; Urquhart & Weir, 1998). Similarly, Alderson (2000) defined reading as the interaction between the reader and the text. The reader looks at print, deciphers the text, determines the meaning of the text, and thinks how the text relates to her background knowledge. Alderson noted that the purpose for reading is a vital part in the reading process, as the reader selects which reading strategies to use depending on why she is reading. If one were reading for pleasure, she would not pay as much attention to specific points in text as she would if she were reading to answer questions on a reading test.

Despite many attempts to describe the act of reading, the definition of reading ability remains elusive. This is because reading is a complex, cognitive process which involves much more than looking at print and assigning meaning to the written letters. Furthermore, the reading process is influenced by an extensive number of variables (Alderson, 2000). These variables can be separated into two categories: reader variables and text variables. Reader variables include the reader's linguistic knowledge, background knowledge, text type knowledge, cultural knowledge, purpose in reading, and motivation. Text variables include text topic, text type, text organization, and text readability. Interaction of all these variables makes the process of reading challenging to understand.

Moreover, the reading process can be different for the same reader on the same text at a different time at a different location with a different purpose in reading (Alderson, 2000). If the process can differ for the same reader depending on the circumstances, it is more likely that the process varies for different readers.

Although researchers have agreed that the nature of reading will never be comprehensively understood (Alderson, 2000; Grabe & Stoller, 2002; Smith, 1988), many have attempted to study the nature of reading by examining the process of reading. For instance, Smith (1971) examined the eye movements of readers while they read a text. By watching how the eyes move, he hoped to understand how the brain internalized text. However, he concluded that what the eye told the brain was not reciprocal to what the brain told the eye. Another way to externalize the reading process was to analyze the mistakes readers made when reading aloud (i.e., miscue analysis) (Goodman, 1969). Despite the effort, it was concluded that the process of reading aloud may be dissimilar to

the process of reading in silence. Other scholars used think-aloud protocols method (e.g., Whitney & Budd, 1996) or immediate recall protocols method (e.g., Bernhardt, 1991) to examine how readers were reading text. Through these studies, researchers found various reading strategies that readers use, and the difficulties they have when processing particular texts. Although these findings have provided interesting insights into understanding the process of reading, it remains challenging to describe the act of reading as it is usually silent, internal, and private (Aebersold & Field, 1997; Alderson, 2000; Crystal, 1997).

Researchers have also attempted to examine the product of reading, which refers to the result of the process. Instead of investigating *how* one reaches the meaning of text, researchers have focused on *what* understanding one reaches after reading (Alderson, 2000). The most common way to examine the product of reading is to administer a reading test and analyze the test results to determine the comprehension of the test-takers. Although this approach is commonly used, there are some limitations. First, the method used to assess one's reading product may affect her test performance. For example, if reading is assessed using a multiple choice format in a community where this test method is uncommon, the test-takers' performance may be affected. This may not be because their reading comprehension level is low, but because they are not familiar with the test format. In this case, the test result would not be an accurate measure of one's reading comprehension. To avoid such results, it is crucial for test developers to learn about reading practices in the community to be tested when designing the test. A second limitation to studying the product of reading is the variation in the product (Alderson, 2000). Because different readers develop different understanding of the text based on

their prior knowledge and experience, it is difficult to determine the "correct" interpretation of the text, if there is such a thing.

Taking the complexity of reading into account, many researchers have designed theoretical models of reading in attempt to depict what happens when one reads (e.g., Goodman, 1976; Gough, 1985; Rumelhart, 1985; Stanovich, 1980). The most commonly discussed reading processing models are: top-down models, bottom-up models, and interactive models. The following sections discuss these models in order to describe reading ability.

2.2.1 Top-down Processing Models

According to the top-down processing model, reading is mostly directed by reader goals and expectations (Grabe & Stoller, 2002). It focuses on the importance of the knowledge that the reader brings to the text. From this perspective, the reader understands and interprets the ideas represented by the text, integrates textual information, links words with their co-referents, makes inferences, forms attitudes about the text and author, and interprets the text as a whole (Grabe & Stoller, 2002; Segalowitz et al., 1991). The reader also fits the text into their prior knowledge (i.e., cultural, syntactic, linguistic, historical) and reconciles the text with their background knowledge when new or unexpected information appears (Aebersold & Field, 1997).

The idea of top-down model was based on schema theory, which accounts for "the acquisition of knowledge and the interpretation of text through the activation of schemata: networks of information stored in the brain which act as filters for incoming information" (Alderson, 2000, p. 17).

One proponent of this kind of model was Goodman (1967), who described reading as a "psycholinguistic guessing game". This metaphor was based on the notion that readers use minimal textual information and maximum background knowledge to guess or predict the author's message from the text. Concurring with Goodman (1967), Smith (1982) stated that readers have the ability to make necessary inferences from their prior knowledge while reading. He further argued that readers read texts selectively and not in a word-by-word manner to eliminate redundancy in the text.

The top-down processing model has had a great impact on ESL reading theory and instruction. For example, Clarke and Silberstein (1977) attempted to transfer Goodman's (1967) idea of psycholinguistic guessing game into practice by developing an instructional framework for ESL reading teachers. They suggested that students need to be taught reading strategies (e.g., guessing meaning from context, drawing inferences from the text, skimming) to become proficient readers. Furthermore, ESL teachers were encouraged to create pre-reading activities to enhance students' reading comprehension, to explain difficult syntax, vocabulary, and organization structure, and assist students to determine strategies for reading.

Although these instructional implications remain valuable to ESL teachers today, the ideas of the top-down processing model have received critical reviews over the years. The metaphor of a psycholinguistic guessing game was criticized as oversimplifying the complex cognitive process of reading. Guessing requires an intricate cognitive process of evaluating the importance of different information in order to reach the best answer (Birch, 2007). However, Goodman (1967) failed to precisely explain how guessing was processed in the reader's brain.

2.2.2 Bottom-up Processing Models

The bottom-up processing model focuses on the information presented by the text (Anderson, 1999). According to this model, the written text is hierarchically organized, and the reader processes text in the following order: looking at the print; recognizing the letter features; decoding them to sound; associating the words to their semantic representations; decoding meanings of words, then phrases, and then sentences (Alderson, 2000; Hedge, 2000; Segalowitz et al., 1991).

Research on readers' eye movements has provided insights into the role of bottom-up processing in fluent reading (e.g., Adams, 1990; Radach, Kennedy, & Rayner, 2004; Rayner & Pollatsek, 1989). Studies have revealed that eyes do not move across a line of text in a continuous manner. Instead, they make a series of pauses called fixations. The reader takes the visual stimulation during the fixations and discards unimportant information when the eyes are changing focus from one point to another (i.e., saccade). During this process, the reader does not guess or sample texts. Instead, they identify the vast majority of words automatically (Grabe, 1991). Furthermore, research on word recognition and lexical access has revealed that readers do not skip large number of words, but process the letters and words thoroughly (Rayner & Pollatsek, 1989). These findings were contrary to what the top-down processing model suggested.

If reading were a guessing game as the top-down processing model indicated, proficient readers and unskilled readers would use guessing skills differently. For example, proficient readers would be more aware of context and constantly refer back to prior knowledge, while unskilled readers would have a difficult time predicting the next words in the sentence. Contrary to this expectation, Perfetti, Goldman, and Hogaboam

(1979) found that unskilled readers are as sensitive to context as proficient readers. Proficient readers automatically and accurately recognize words so that there is no need for guessing. Proficient readers identify the word then move quickly to a higher level of prediction and monitoring during the fixation, while unskilled readers take longer to recognize the word at each fixation. In other words, what distinguishes skilled readers from unskilled readers is "not the number of letters in a fixation, nor the number of words fixated per page, but the speed of the fixation – the automaticity of word recognition – and the processes that occur during fixation" (Alderson, 2000, p. 18). In summary, research has shown that automatic and precise word recognition ability appears to be a vital component of reading ability (Adams, 1990; Stanovich, 1986).

2.2.3 Interactive Models

Both top-down and bottom-up processing models were helpful in conceptualizing the reading process, but neither of them were an adequate characterization of the reading process as they "unhelpfully polarize a description of how mental processes interact with text features in fluent reading comprehension" (Day & Bamford, 1998, p. 12). The most comprehensive description of the reading process model is known as the "interactive model" (Rumelhart, 1985), which combines the elements of top-down and bottom-up models. Perhaps the best known advocate of interactive models is Stanovich (1980, 1986, 2000), who calls his model an 'interactive-compensatory' model. This model is developed from cognitive psychology using the ideas of LaBerge and Samuels (1974) and Perfetti and Lesgold (1977, 1979).

Stanovich (2000) describes the term 'compensatory assumption' as "the assumption that deficiencies at any level in the processing hierarchy can be compensated for by a greater use of information from other levels, and that this compensation takes place irrespective of the level of the deficient process (p. 49). In other words, the compensatory component refers to the idea that a weakness in one area of knowledge can be compensated for by strength in another area (Urquhart & Weir, 1998). For example, poor readers who have weak word recognition skills might rely more on contextual clues (Stanovich, 1980).

The interactive-compensatory model is composed of two contextual mechanisms. One is "an automatic spreading activation process operating in semantic memory" (Stanovich, 2000, p. 50). This mechanism is often used by good readers who automatically process text and access little cognitive capacity. Good readers attempt to understand the meaning of the text as a whole without paying too much attention to any particular word. They have good decoding skills so that they do not rely much on context information.

The other mechanism is "a process of specific contextual prediction that operates more slowly, utilizes attentional capacity, and causes facilitation (Stanovich, 2000, p. 50). This mechanism is often employed by poor readers who attempt to make sense of reading by heavily relying on context. The notion of poor readers using contextual features more than the good readers has been empirically supported (e.g., Becker, 1982; Briggs, Austin, & Underwood, 1984).

In summary, according to the interactive compensatory model, readers use both lower-level and higher-level processing skills. Lower-level skills include skills such as

rapid, automatic, linguistic processing so that the reader can recognize words, decode unfamiliar words, and examine the part of speech of a particular word. In doing so, basic grammatical information can be obtained to grasp clause-level meaning. Higher-level skills include skills such as comprehension and interpretation so that the reader can anticipate what happens next in the text and draw on past experiences. Researchers have come to consensus that both bottom-up and top-down processes interact either simultaneously or alternately in fluent reading (Aebersold & Field, 1997; Anderson, 1999; Grabe, 1991; Grabe & Stoller, 2002; Murtagh, 1989). The balance between low and high processing skills is likely to vary with text, reader, and purpose (Alderson, 2000); nevertheless, both processing skills are crucial in reading. Therefore, the current study incorporates the idea of an interactive model.

2.2.4 Testing Skills of Reading

Just as there have been numerous theories proposed for determining how reading is processed, many researchers (e.g., Grabe, 1991; Lunzer & Gardner, 1979; Munby, 1978) have attempted to define what it means to be able to read by identifying skills of reading. Skills of reading, according to Alderson (2000), are "the notion that the act of reading consists of the deployment of a range of separate skills, abilities or strategies" (p. 93). A list of skills such as the one Munby (1978) composed, is helpful in diagnosing a reader's problems or creating test tasks/items. However, many of the proposed taxonomies (e.g., Bloom, Engelhart, Furst, Hill, & Kratwohl, 1956; Munby, 1978) were not built based on empirical evidence. Moreover, many of the postulated skills overlap in terms of definition, which makes it difficult for experts to agree on what skills are being

identified or tested by which test item (Alderson, 2000). Despite the criticisms, identifying skills remains helpful and practical in testing situations, and cannot be overlooked when examining the nature of reading.

There have been numerous studies of reading skills. For example, Lennon (1962) examined various factor analytic studies on reading skills. He identified the components of reading ability as: a general verbal factor, understanding of explicitly stated matter in text, and understanding of implied meaning in the text. Similarly, Carroll (1993) reviewed numerous factor analytic research studies on reading and concluded that there are mainly four common factors in reading: (1) general reading comprehension, (2) special (as opposed to general) reading comprehension, (3) reading decoding, and (4) reading speed. From a pedagogical point of view, Anderson (1999) stated that "understanding main ideas, making inferences, predicting outcomes, and guessing vocabulary from context are all reading skills that readers of English typically need to develop" (p.1). In other words, these skills are what ESL learners need to acquire when learning to read. Readers use both top-down and bottom-up processing skills to grasp the main ideas of the passage, understand the details of the text, make inferences, and speculate about the meaning of unknown words based on the context.

2.2.5 Model for the Current Study: Measuring Reading Ability

As we have seen in the previous section, second language reading ability is a multifaceted, complex construct, which involves both lower- and higher-level processing. In order to operationalize the test based on the literature review, and to identify what skills are measured by a reading test, researchers must first identify the purpose of the test

they are using for their study. According to the English Language Institute of University of Michigan (2006c), the ECPE reading test measures the test-takers' ability to: (1) recognize the main idea of the text; (2) recognize the relationship of ideas within a text; (3) recognize the author's organizational pattern and method of argument; (4) synthesize information, both main ideas and specific details; (5) recognize and understand supporting details and examples, and their functions; (6) recognize and understand specific lexical items in context; (7) understand referents; (8) draw inferences and conclusions based on the passage; (9) distinguish between fact and opinion; (10) understand the author's attitude and/or opinion; and (11) understand the author's purpose (p. 8). Because some of these skills share similar characteristics (e.g., recognize the main idea and synthesize the main idea), these eleven types of abilities claimed by the test developer are grouped together to measure four summary components that are arguably more distinct: reading for the main idea, vocabulary in context, inferences, and detail.

There are three main reasons for grouping the similar types of reading skills. First, there are only twenty reading items per test and not every test measures all eleven specific skills that test developers' postulated. Hence it is not possible to compare test forms if each test measures a different subset of these detailed skills.

Second, some of the eleven skills share similar characteristics and it would be difficult for the judges to agree on what skills are operationalized by which item (Alderson, 2000). Therefore, it was necessary for the skills to be categorized into more aggregated and distinct groups where disagreements should be minimal.

Third, according to Grabe (1999), reading assessment is typically driven by assessment theory and the reasonably strong psychometric qualities of traditional reading

comprehension tests. He further stated that "simple and straightforward measures of main idea and detailed comprehension questions on passages, combined with sections on vocabulary, provide strong reliability and at least arguable validity for these testing approaches" (Grabe, 1999, p. 35). In other words, testing for main ideas, details, and vocabulary generally provide reasonable reliability and validity.

Based on these reasons, the four reading components were operationalized for the current study. The first component, main idea, is measured based on the ability to identify what the author aims to express as the gist or the most relevant idea. Out of eleven abilities the test developers listed, the following two are measured under the main idea component: (1) recognize the main idea of the text, and (4) synthesize information, both main ideas. Example main idea items are: *What is the main idea of this passage?* or *The main idea of this passage is to...*

The ability to understand detail is measured based on the ability to recognize specific information explicitly stated in the text. The following ability is measured under detail: (5) recognize and understand supporting details and examples, and their functions. Example detail items are: *According to the passage, which factor is most important in...?* or *Where do the lions sleep at night?*, if the place the lions sleep is explicitly mentioned in the text.

Vocabulary in context is measured based on the ability to understand the meaning of words in the provided context. The ability to (6) recognize and understand specific lexical items in context, is measured under this component. An example vocabulary item is: *What does the word "bank" in the third sentence mean?*

The ability to make inferences is measured based on the ability to draw logical conclusions about the author's intentions or attitudes in the text, identify the genre, understand the cohesion and coherence in the text. Out of eleven abilities listed, seven are categorized as inference: (2) recognize the relationship of ideas within a text (i.e., coherence), (3) recognize the author's organizational pattern and method of argument (i.e., genre identification), (7) understand referents (i.e., cohesion); (8) draw inferences and conclusions based on the passage; (9) distinguish between fact and opinion; (10) understand author's attitude and/or opinion; and (11) understand author's purpose. Examples of inference items are: *What is the tone of the author?* or *Where would the lions go to find food?* Even if the place the lions go to look for food is not explicitly stated in the text, but the answer can be derived implicitly (inferred) from the text or even outside the text.

In summary, the current study used four reading components (i.e., main idea, detail, vocabulary in context, and inference) to operationalize reading ability measured in the ECPE reading section. Table 2.3 presents how the test-takers' ability is grouped together for the purpose of this study. It also provides example items for each variable.

Table 2.3
Ability Measured in the Reading Section of ECPE

	Test-Takers' Ability to:	Example Items
Main Ideas	• recognize the main idea of the text • synthesize information main ideas	• The main purpose of this passage is to… • What is the main idea of this passage?
Details	• recognize and understand supporting details and examples, and their functions	• According to the passage, which factor is most important in …? • Where do the lions sleep at night?
Vocabulary in Context	• recognize and understand specific lexical items in context	• What does the word "bank" in the third sentence mean?
Inferences	• recognize the relationship of ideas within a text (i.e., coherence) • recognize the author's organizational pattern and method of argument (i.e., genre identification) • understand referents (i.e., cohesion) • draw inferences and conclusions based on the passage • distinguish between fact and opinion • understand author's attitude and/or opinion • understand author's purpose	• What is the tone of the author? • What is the author's purpose of this article?

The next section focuses on the studies using a structural equation modeling approach in the field of language assessment. First, it focuses on reviewing SEM studies using a single group of participants. Then, it focuses on SEM studies using multiple groups of participants. The multi-group studies provide examples of how SEM can be a useful tool in comparing the underlying trait structures of two test forms.

2.3 SEM Studies: Single-Group Studies

Structural equation modeling (SEM) was first employed in a study of language testing by Bachman and Palmer (1981a). Although several researchers have employed the

technique in their studies (e.g., Bachman & Palmer, 1981b, 1982, 1989; Fouly, 1985; Purcell, 1983; Turner, 1989), it remains an unfamiliar statistical procedure to many scholars. Kunnan (1998) listed a number of language testing studies using SEM, and stated that "this short list illustrates that SEM applications in language assessment research have been very few in number and the range of investigations equally small" (p. 297). The number of L2 assessment studies using SEM, however, has dramatically increased as researchers have recognized the usefulness of this statistical procedure. The following sections describe how SEM has been used in the field of language testing.

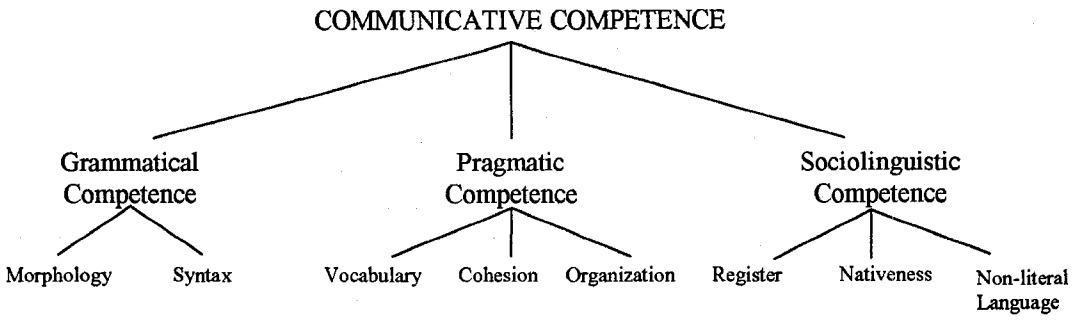
2.3.1 SEM Studies on the Nature of L2 Proficiency

Language testing researchers initially used SEM to examine construct validity of L2 proficiency. Bachman and Palmer (1981b, 1982, 1989) conducted a series of construct validation studies using SEM to investigate the factor structure of language proficiency. In their first study (1981b), they examined the construct validity of the Foreign Service Institute (FSI) oral interview using a multi-trait multi-method (MTMM) matrix. The matrix was comprised of six measures representing combinations of two traits (speaking and reading) and three methods (interview, translation, self-rating). The participants of the study were 75 native Mandarin-Chinese-speaking learners of English. The data were analyzed using confirmatory factor analysis (CFA) to examine the extent to which relationships predicted by theoretical models fit the data. They tested four theoretical models, which correspond to different hypotheses regarding the nature of language proficiency. The models depicted L2 proficiency as (1) completely divisible (i.e., two distinct, uncorrelated trait factors), (2) partly divisible, (i.e., two distinct, but correlated

trait factors), (3) somewhat divisible (i.e., a general factor plus two uncorrelated trait factors), and (4) completely unitary (i.e., a single trait factor). Based on chi-squares, probabilities, degrees of freedom, and the loadings of each measure on the various factors, they determined that both the partly divisible model and the somewhat divisible model adequately fit the data. However, the partly divisible model was more parsimonious and substantively more meaningful. Hence, they concluded that the construct measured by the FSI oral interview was partly divisible with two distinct, but correlated factors. In other words, speaking and reading represented distinct, yet correlated traits.

Bachman and Palmer (1982) conducted another study investigating the construct validation of a test which included an oral interview, a writing test, a multiple-choice grammar test, and a self-rating questionnaire. With this test, they examined whether the hypothesized theoretical framework of communicative language ability was empirically supported. The theoretical framework under investigation comprised three traits: grammatical competence, pragmatic competence, and sociolinguistic competence. Grammatical competence included morphology and syntax. Pragmatic competence encompassed vocabulary, cohesion, and organization. Sociolinguistic competence consisted of register, nativeness, and non-literal language. The model is shown in Figure 2.5.

Figure 2.5
Model of Communicative Language Ability
(Adopted from Bachman & Palmer, 1989, p. 17)



The participants of the study were 116 non-native speakers of English from 36 different countries. Bachman and Palmer (1982) used an MTMM matrix, which comprised twelve measures representing combinations of three traits (grammatical competence, pragmatic competence, sociolinguistic competence) and four methods (oral interview, writing test, multiple-choice test, and self-rating questionnaire). The data were examined using CFA to evaluate the extent to which the postulated model fit the observed data.

The results indicated that there were two distinct factors. Contrary to their postulated framework, grammatical competence and pragmatic competence loaded on one factor while sociolinguistic competence loaded on another. Based on the results, Bachman and Palmer speculated that grammar and vocabulary were necessary for cohesion and organization, and that these were functions of the organizational aspects of language. On the other hand, sociolinguistic competence related more to the affective aspect of language.

Bachman and Palmer (1989) conducted another study on construct validation. This time, they focused on the construct validity of a self-rating questionnaire of

communicative language ability. They used the same questionnaire data as the previous study (Bachman & Palmer, 1982); hence, the participants and the questions on the questionnaire remained the same. The previous study had shown that the questionnaire was reliable and provided information on general language ability. With the 1989 study, they attempted to determine whether different question types would provide evidence that the questions were reliable and valid. They also investigated the relationship between question types and the traits these questions were designed to measure.

Bachman and Palmer (1989) used the same communicative language ability model as the 1982 study to determine whether the questionnaire adequately measured the postulated traits. To analyze the data, they again used an MTMM matrix, which comprised of nine measures representing combinations of three traits (grammatical competence, pragmatic competence, sociolinguistic competence) and three question types ('ability' questions, 'difficulty with production' questions, and 'recognition' question). The data were examined using CFA to evaluate the extent to which the postulated model fit the observed data.

The results indicated that self-ratings can be reliable and valid measures of communicative language ability (Bachman & Palmer 1989). The reliability estimates were high and the self-rating measures loaded on a general factor. They also found that 'difficulty with production' questions on grammar and sociolinguistics had reasonably high loadings on their respective trait factors. This meant that the self-rating questions on grammar difficulty and sociolinguistic difficulty were good measures of the participants' communicative language ability.

Bachman and Palmer (1989) compared their study findings to that of Davidson and Henning's (1985) study on self-rating. Davidson and Henning had contradictory findings on self-rating questionnaire, which indicated that a self-rating questionnaire was not a good indicator of language ability. Bachman and Palmer provided possible reasons why the findings were different. One of the reasons was the difference in the statistical procedure used in the studies. Bachman and Palmer used CFA while Davidson and Henning used IRT. Bachman and Palmer stated that factor analysis is "appropriate for examining the relationship of groups of items to many different underlying factors, whereas the IRT modeling used by Davidson and Henning is best suited to examining the extent to which individual items fit a single underlying dimension" (1989, p. 23). Bachman and Palmer (1989) acknowledged that complementary roles of IRT and factor analysis by stating that IRT is a useful tool for test development while factor analysis is one approach to test validation.

2.3.2 Studies on Test-takers' Cognitive Abilities and Strategies

The use of SEM has not been limited to studying the nature of L2 proficiency. Researchers have also employed SEM to examine the relationship between L2 proficiency and test-taker's cognitive abilities. Sasaki (1993) investigated the relationships among foreign language aptitude, intelligence, and second language proficiency (SLP) using SEM. She first investigated the trait structure of SLP, and then examined the relationship between the proposed general SLP factor and a general cognitive factor that was assumed to influence foreign language aptitude and intelligence. The results indicated that the general SLP factor and the general cognitive ability factor

were two distinct, but mutually correlated factors. Moreover, aptitude was the best indicator of the general cognitive ability factor.

Purpura (1997) examined the factor structure of test-takers' cognitive and metacognitive strategy use and second language test performance. The 1382 participants took an 80-item strategy questionnaire, followed by a 70-item L2 language test. Purpura first separately examined the trait structure of second language test performance, cognitive strategy use, and metacognitive strategy use. Then, he examined the relationships between strategy use and second language test performance using a full latent variable model. The results indicated that second language test performance was measured by two factors: lexico-grammatical knowledge and reading ability. He found that lexico-grammatical knowledge and reading ability were not only correlated, but lexico-grammatical knowledge predicted reading ability. As for cognitive strategy use, it was directly and positively associated with second language test performance. Metacognitive strategy use had a direct and positive relationship with cognitive strategy use and was indirectly related to second language test performance. Metacognitive strategy use appeared to exert an executive function over cognitive strategy use. This study provided a more comprehensive view of cognition and second language test performance to the field of language assessment.

Another strategy use study using SEM was conducted by Xi (2005). She examined how task characteristics and test-taker characteristics influenced the strategies used in a semi-direct oral test. The oral test she examined contained a graph description task, which required the test-takers to view the graph and describe it in a given time. The test-taker's performance on the graph task was determined not only by his or her

speaking ability, but also by the comprehension of the graph, the graph familiarity, and the characteristics of the graph. To understand the nature of a graph task, she investigated the relationships among task characteristics (the number of visual chunks and the amount of planning time), test-taker characteristics (graph familiarity and general speaking proficiency), and test performance (holistic scores on the experimental graph tasks). The participants were 236 international graduate students from 34 native language backgrounds studying in the U.S. The oral test used for this study was the SPEAK exam developed by Educational Testing Service. Xi used SEM to model the relationships among these factors and examined how the hypothesized models fit the observed data. The results indicated that the test-takers' graph familiarity influenced the overall communicative ability. Hence, she suggested that test-takers' graph familiarity may be a potential source of construct-irrelevant variance. The results also showed that reducing the number of visual chunks in a graph and allowing planning time to test-takers positively impacted the cognitive processes involved in graph comprehension. Furthermore, limiting the number of visual chunks and providing planning time helped lessen the influence of graph familiarity. These findings empirically demonstrated the need to provide planning time to test-takers with a graph task in oral test.

A separate study focused on cognitive and affective factors in writing was conducted by Lee (2005). He investigated how cognitive/affective factors and reading/writing behavior interacted with one another. He further examined how these factors influenced test-takers' writing performance. The factors he examined were: self-initiated reading, self-initiated writing, writing apprehension, writer's block, and attitudes toward instruction. He first tested the relationship among these factors using confirmatory

factor analysis. Then, he used a full structural model to investigate how these factors impacted test-takers' writing performance. The participants were 270 university students taking an English writing course in Taiwan. Their native language was Mandarin Chinese. The ability level of the participants ranged from low to high intermediate. The participants were given a composition test and three questionnaires. The questionnaires were: the writing apprehension scale (Daly & Miller, 1975a, 1975b), the writer's block questionnaire (Rose, 1984), and a questionnaire asking participants' involvement in and attitudes toward different literacy activities. The findings showed that participants did more self-initiated reading than writing, and the more reading one did, the more one would engage in writing. The results also indicated that self-initiated reading helped reduce writer's block. Writer's block and writing apprehension were interrelated, but neither factor was associated with writing performance. The test-takers who reported high writer's block did not necessarily perform poorly on the composition test. As for attitude toward instruction, it failed to predict writing apprehension and writer's block. It also failed to predict writing performance. In other words, the participants' attitudes toward instruction had nothing to do with how well they performed on the composition test. Lastly, self-initiated reading was the only significant predictor of writing performance. The more the test-taker read on his/her own, the higher was the score on the composition test.

Lee (2005) provided valuable insights into the comprehensive understanding of EFL writing. However, he may have obtained different results if he had performed multi-group SEM. The participants in the study represented a fairly wide range of English language proficiency levels; hence, he could have separated the low-ability group from

high-ability and performed multi-group SEM. It would be interesting to investigate whether the model of writing is invariant across low-ability group and high-ability group.

2.4 SEM Studies: Multi-Group Studies

Several studies have focused on SEM applications involving more than one sample group. This type of application is called multi-group SEM. There are two types of multi-group analyses: non-simultaneous multi-group and simultaneous multi-group analyses. The former type separately compares the groups while the latter type simultaneously compares the groups. The distinction is discussed further with the provided example studies (Bae & Bachman, 1998; Purpura, 1998 for the simultaneous multi-group example studies, and Kunnan, 1995 for a non-simultaneous multi-group example study).

The central concern of multi-group SEM is to explore whether components of the factor structures are invariant across particular groups. Many studies in the field of language testing have applied multi-group SEM (e.g., Bae & Bachman, 1998; Ginther & Stevens, 1998; Kunnan, 1995; Purpura, 1998, 1999; Pyo, 2001; Yun, 2005). Researchers often divided the groups into either different ability level groups (e.g., Bae & Bachman, 1998) or different language background groups (e.g., Kunnan, 1995). The following sections describe how researchers have used multi-group SEM in the field of language testing.

2.4.1 Comparing Different Ability Level Groups

Bae and Bachman (1998) were one of the early researchers to conduct a study using simultaneous multi-group covariance structure analyses. They investigated the equivalence of factor models of reading and listening abilities across two groups. One group was Korean American (KA) students (N=120) whose first language was Korean, and the other group was non-Korean American (non-KA) students (N=36) whose primary language was English. Both groups were elementary school children who attended a Korean/English two-way immersion program. The students in this program learn both English and Korean; however, the study focused on the participants' reading and listening skills in Korean. The participants took a listening test with three tasks, and a reading test with three tasks.

Bae and Bachman (1998) first investigated the underlying trait structures of reading and listening skills for each group (i.e., baseline model) using confirmatory factor analysis (CFA). They then performed simultaneous covariance structure analysis to examine whether the hypothesized model of reading and listening was equivalent across KA and non-KA groups. Once the baseline model was established for each group, individual parameters were compared using equality constraints across groups.

The results of CFA indicated that the listening and reading were distinct factors. The simultaneous covariance structure analysis showed that KA and non-KA groups had the same underlying factor model of reading and listening. Furthermore, the correlation between the listening and reading was high for both groups. Factor loadings for all tasks were essentially the same across the two groups except for one listening task. This listening task may have had a task-specific effect, which interacted with individual

characteristics of the two types of learners (Bae & Bachman, 1998). All the findings were theoretically supported; however, the small sample size (N=36) of non-KA group may have impacted the results.

Similar to Bae and Bachman (1998), Purpura (1998) performed a simultaneous multi-group covariance structure analysis to investigate how the model of strategy use and second language test performance (SLTP) differed across low-ability and high-ability groups. The total participants were 1,382 EFL students who took two types of strategy questionnaires and an English proficiency test.

Prior to performing multi-group analyses, Purpura (1998) investigated the model of strategy use and SLTP for all the participants (i.e., single-group analyses). Subsequently, he separately investigated the model for both low-ability group (N=941) and high-ability group (N=234). Then, he performed simultaneous multi-group analyses of the relationships between strategy use and SLTP to determine the degree to which these models were invariant across the two groups.

The results showed that there were both similarities and differences between the low-ability and high-ability groups in how strategy use impacted their performance. Both groups produced an almost identical underlying factorial structure for metacognitive strategy use and SLTP, whereas cognitive strategy use produced different models for each group (Purpura, 1998). The high-ability and low-ability test-takers used strategies differently on language tests. This study has shown the importance of separately investigating the models of strategy use and SLTP as the model can vary depending on the ability level.

Another study comparing a low-ability and a high-ability group was conducted by Yun (2005). She investigated the relationship between L2 writing test performance and its explanatory variables among Korean EFL learners. She used a simultaneous multi-group SEM to determine whether the factorial structures of L2 writing were invariant across the low- and high-ability groups. The hypothesized writing model for this study consisted of five variables: L1 writing ability, L2 language knowledge, L2 writing experience, L2 reading experience, and test preparedness. To investigate whether the model of writing ability was comparable for different ability groups, she first examined the baseline model for the entire sample. Then, she divided the participants into a high-ability group (N=153) and a low-ability group (N=147) and simultaneously modeled the two groups by imposing equality constraints across groups. The participants took a writing test both in English (L2) and Korean (L1), they answered a multiple-choice English cloze test, and completed a questionnaire which asked background information concerning the amount of experience they had in L2 writing and reading. The writing tests were scored using both holistic and analytic rating scales.

The results indicated that writing ability in L1 and L2 appeared to be two distinct yet correlated factors. The hypothesized writing model fit well with L2 language knowledge being the primary predictor for L2 writing performance. The results of the simultaneous multi-group SEM showed that the relative importance of the five variables included in the model was not the same for both the low- and high-ability groups. For the high-ability group, the L2 writing experience variable was insignificant. For the low-ability group, the L1 writing ability variable was insignificant. The L2 language knowledge variable was the primary predictor for both groups; however, the effect in

explaining the variance in L2 writing performance was smaller for the high-ability group (Yun, 2005). She, therefore, concluded that these groups should be treated as coming from different populations. Similar to Purpura's (1999) study, this study has shown that it is important to consider the ability level of the participants because the model in question may be variant among the participants.

2.4.2 Comparing Different Language Groups

Multi-group studies were not limited to comparing groups with different ability levels. Some researchers have utilized multi-group SEM to compare different language groups. For instance, Kunnan (1995) explored the effect of the test-takers' background characteristics on EFL test performance across Indo-and non-Indo-European language groups. The participants (N=985) were mainly EFL students in eight countries (Thailand, Egypt, Japan, Hong Kong, Spain, Brazil, France, and Switzerland). Kunnan categorized the participants from Thailand, Egypt, Japan, and Hong Kong as a non-Indo-European group (N=380) and Spain, Brazil, France, and Switzerland as an Indo-European group (N=605). All participants took a background questionnaire and a series of standardized English proficiency tests.

Kunnan (1995) first performed an exploratory factor analysis on the entire population as an initial exploration of the data. Then, he separately modeled test-taker characteristics and test performance. Based on the results, a four-factor model of test-taker characteristics and a four-factor model of test performance emerged. He then modeled the relationship between test-taker characteristics and test performance for both Indo-and non-Indo-European language groups. When comparing the two groups, he used

a method called non-simultaneous multi-group analysis. With this method, a single-group analysis for each group was separately modeled. Then, the model of test-taker characteristics and test performance for Indo-and non-Indo-European groups were compared. The results indicated that all test-taker characteristics factors (i.e., information exposure, instruction, location of exposure, and monitoring) had influence on proficiency test performance. The amount of influence, however, differed depending on the language group.

Kunnan's (1995) study compared the factor models of different language groups using non-simultaneous analysis. Although he succeeded in providing substantive and interpretable models for the two groups, he failed to provide analytical comparisons of the invariance of the factor loadings, error variances, and factor correlations (Stricker, Rock, & Lee, 2005). Furthermore, the use of simultaneous multi-group analysis (i.e., testing invariance of the model of each group by simultaneously estimating the parameters for both groups) might have provided more precise information about the models of test-taker characteristics and test performance for the two language groups.

2.4.3 Summary of Multi-Group SEM Studies

All of the multi-group SEM studies discussed above have focused on comparing either different ability level groups (e.g., Bae & Bachman, 1998; Purpura, 1999; Yun, 2005) or different language groups of test-takers (e.g., Kunnan, 1995; Ginther & Stevens, 1998). Although there has been no research comparing two different test forms using multi-group SEM in the language testing literature, the concept is the same as comparing two different groups of a population. The multi-group SEM studies have shown that it is

possible to compare two different populations using SEM. Therefore, comparing two different test forms using SEM may be feasible. Based on this concept, the present study used multi-group SEM to compare the factorial structure of two test forms.

2.5 Summary

The first part of the chapter focused on defining the theoretical construct of the GCVR section. It reviewed how lexico-grammatical knowledge and reading ability have been defined and conceptualized in the field of language teaching and assessment. Then, the purpose of the GCVR sections was reviewed. By reviewing both the substantive theories and purpose of the GCVR section of ECPE, the hypothesized underlying trait structures of GCVR section was developed. The second part of the chapter reviewed the studies using SEM, which demonstrated SEM as an efficient method of analysis for research in the field of language testing. In addition, the application of SEM in the current study contributes to providing construct validity evidence of the GCVR section of ECPE. The next chapter addresses the research design and method used in the current study to investigate the research questions.

Chapter III

METHODOLOGY

This chapter describes the methodological procedures that were used in the current study. It provides an overview of the design, the participants, and the measurement instruments used in the current study. Procedures related to data collection and data coding are explained, study variables are operationalized, and the steps in data analyses are described. The chapter concludes with the description of the hypothesized models used for the current study.

3.1 Design

There are two main parts to the study. The first part investigates the underlying trait structures of the grammar/cloze/vocabulary/reading (GCVR) section of the ECPE. The purpose of this part of the study is to provide evidence that the GCVR section measures what it is intended to measure (i.e., lexico-grammatical knowledge and reading ability). The second part of the study focuses on the comparability of the underlying trait structures across two different test forms of ECPE.

3.2 Participants

The data were collected by the authorized test centers of the English Language Institute of the University of Michigan (ELI-UM). The test was administered at over 125 test centers in 20 countries in 2003-04 (N=33,662) and 2004-05 (N=32,473). Between the

two test administrations, it involved a total of 66,135 test-takers of English as a foreign language (EFL). The native languages of the participants fell into six different categories: Afro-Asian, Austronesian, African and Transafrican, Eurasian, Indo-European, and Sino-Indian. Greek was categorized separately from the other Indo-European languages because the majority of participants were Greek speakers (over 90 percent) in both administrations. Other Indo-European language speakers accounted for 8 percent of the test-takers. The remaining language groups accounted for less than 1 percent of the total number of test-takers. The breakdown of participants by their native language is shown in Table 3.1.

Table 3.1
Native Language of Participants
(Adapted from English Language Institute, 2006b, 2006c)

Language Background	2003-04		2004-2005	
	# of participants	%	# of participants	%
Afro-Asian	80	0.24	133	0.41
Austronesian	30	0.09	10	0.03
African/Trans African	7	0.02	3	0.01
Eurasian	70	0.20	56	0.17
Greek	30753	91.36	29641	91.28
Other Indo-European	2612	7.76	2549	7.85
Sino-Indian	3	0.01	3	0.01
Missing (none listed)	107	0.32	78	0.24
Total	33662	100.00	32473	100.00

Participants were asked to provide their date of birth on the test registration form. For the 2003-04 administration, the mean age was 21 and the median was 20. There were two high frequency age groups: Age 13-16 (30.26%) and Age 20-22 (24.35%). The mean age was 21.92 and the median was 20 for the 2004-05 administration. The age distribution again showed two high frequency age groups: Age 13-16 (22.81%) and Age 20-22 (24.61%). However, the percentages for the age group 13-16 were noticeably

different. There were about 7.5 percent more test-takers in the age group 13-16 in the 2003-04 administration. In general, the test-takers in the 2004-05 administration were somewhat older than the test-takers in the 2003-04 administration. The age distribution of the participants for both administrations is shown in Table 3.2. Based on the native language and age information of the participants, the two test administrations appeared to have, not identical, but a similar population.

Table 3.2
Age Distribution of Participants
(Adapted from English Language Institute, 2006b, 2006c)

Age	2003-04		2004-05	
	# of participants	%	# of participants	%
12 and less	40	0.12	7	0.02
13-16	10186	30.26	7407	22.81
17-19	4625	13.74	5296	16.31
20-22	8197	24.35	7992	24.61
23-25	4369	12.97	4971	15.31
26-29	3191	9.48	3283	10.11
30-39	2467	7.33	2802	8.63
40 and more	587	1.74	715	2.20
Total	33662	100.00	32473	100.00
Mean		21.10		21.92
Median		20.00		21.00
Standard Deviation		6.73		6.63

3.3 Measurement Instruments: The ECPE Test

Developed by the English Language Institute of The University of Michigan (ELI-UM) for advanced-level students in 1953, the ECPE is designed to measure the test-takers' English language performance levels in all skill areas (speaking, listening, writing, and reading) of language use (English Language Institute, 2006a). The content and difficulty of the test are intended to reflect the English language skills required of a university level student.

The ECPE consists of two types of tests: the preliminary test and the final test. The preliminary test is used as a screening device for the final test. This test is not obligatory, but recommended to the people who are interested in taking the final ECPE. It helps them become familiar with the contents of the exam and provides a good estimate on how they would be expected to perform on the final test. There are 35 grammar, vocabulary, MC cloze, and reading items on the preliminary test, which take 30 minutes to complete. All items are in a selected-response format.[3]

The final ECPE contains four sections: speaking, writing, listening, and grammar/cloze/vocabulary/reading (GCVR). A certificate of proficiency is issued to the test-takers who pass all four sections of the test, and those with high scores in all four sections receive a certificate with honors. This certificate is recognized in many countries as evidence of advanced proficiency in the English language for education, employment, career advancement, and business purposes (English Language Institute, 2006a).

3.3.1 Description of the Each Section in the ECPE Test

There are four sections in the ECPE test: speaking, writing, listening, and grammar/cloze/vocabulary/reading (GCVR). The speaking task is a 10 to 15 minute one-on-one oral interview, which contains both short exchanges and longer discourse. For the writing task, test-takers are asked to write an essay on one of two assigned topics in 30 minutes.

The listening and GCVR sections are selected-response type questions. The listening section consists of 50 items, which is designed to measure the test-takers' aural English ability at the advanced level (English Language Institute, 2006c). These items are

[3] The preliminary test is not analyzed in the current study.

intended to assess test-takers' ability to comprehend both direct and indirect meaning of a conversation or an extended monologue. The test time for the listening section is 35 to 40 minutes.

The GCVR section consists of 120 multiple-choice items with four separate parts: grammar (40 items), cloze (20 items), vocabulary (40 items), and reading (20 items). The test time for the GCVR section is 75 minutes. The grammar items are intended to measure the test-takers' ability to recognize appropriate grammatical forms to convey and interpret direct and indirect meaning to a given context (English Language Institute, 2006c). The cloze items are designed to measure the test-takers' ability to read and comprehend texts, and to choose an appropriate word to fill in a gap. The vocabulary items are designed to assess knowledge of words frequently used in academic or business discourse (English Language Institute, 2006c). The reading items are intended to measure the test-takers' ability to understand a university-level reading text. Table 3.3 summarizes the sections of the ECPE test format.

Table 3.3
Description of the Current ECPE Test Format

Section	Tasks	Time (minutes)	Number of Items
1	Speaking (interview)	15	1 task
2	Writing an Essay	30	1 task
3	Multiple-Choice Listening	35-40	50
4	Multiple-Choice (GCVR)	75	
	Grammar		40
	Cloze (1 passage)		20
	Vocabulary		40
	Reading (4 passages)		20

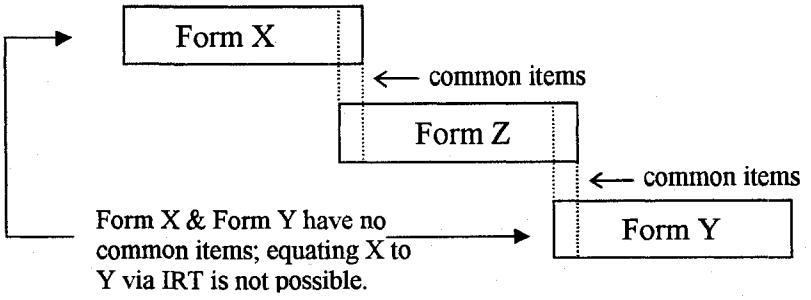
3.3.2 ECPE Form Used in This Study

The two test forms used in the current study were administered in 2003-04 and 2004-05. The two test forms are called Form X (2003-04) and Form Y (2004-05) in the rest of this study.

Although the GCVR section consists of 120 items, 20 items (10 grammar items and 10 vocabulary items) were subtracted from the data analysis in this study. This is because these items were trial items and were not used for scoring. Therefore, a total of 100 items was used in this study: grammar (30 items), MC cloze (20 items), vocabulary (30 items), and reading (20 items).

It is customary for the test developers at the University of Michigan to include a set of common items in ECPE tests to equate test forms. The two ECPE test forms used in this study were equated by linking the scores to yet other ECPE test forms which shared items in common with X and Y individually, but not jointly. In other words, Form X is equated with a third form, Form Z, by including common items in the two forms. Similarly, Form Y is equated with Form Z by including different set of common items in the two forms. In theory, Form X and Form Y are equated, though they do not share common items in these forms. This is illustrated in Figure 3.1. The forms used in the current study are Form X and Form Y with no common items; therefore, it is not possible to perform item response theory to equate test forms.

Figure 3.1
Test Equating and ECPE Forms



3.4 Procedures

3.4.1 Administration of Instruments

The ECPE is administered annually from November to April, depending on testing location. There are over 125 testing centers in about 20 countries. All certified testing centers must meet the standards required by the ELI-UM: (a) the test rooms must be reasonably comfortable with minimal distractions, (b) the test rooms must be large enough to assign test-takers to sit in alternate seats, (c) there must be one proctor assigned for every 25 to 35 test-takers. Test centers are required to secure the test during exam delivery, storage, and administration so that no test item information is passed to the test-takers prior to the test date.

On the day of the test, each test-taker is asked to present two pieces of identification to ensure the integrity of the test scores. During the exam, no questions regarding the test items are answered.

The writing, listening, GCVR sections are given during a single administration period in that order. For the writing section, test-takers are asked to write their response in thirty minutes to one of two given prompts provided on the test. At the end of this section, the administrators collect the essay papers and proceed to the listening section.

The administrators start the tape and the test-takers follow the directions and answer each selective-response item. The listening section takes about 35 to 40 minutes. The last section is GCVR. The test-takers are asked to read the directions and fill in the small circles to indicate an answer for each item on the provided computer-scannable bubble sheet. The GCVR section is administered in 75 minutes. Upon completion of the test, the answer sheets and test booklets are collected.

The speaking task called, Interactive oral communication section, is scheduled on a different date. This is a 15 minute face-to-face interaction with an examiner and the conversation is recorded on a tape.

After both test administrations are completed, the answer sheets, essay papers, interview tapes, listening section tapes, and all other relevant paperwork are returned to the ELI-UM. The test booklets and other secure test materials are destroyed at the test centers.

3.4.2 Scoring

The preliminary test consists of 35 multiple-choice items and is scored dichotomously based on right or wrong responses. The maximum score on the preliminary test is 35. ELI-UM suggests that test-takers who score 23 or higher on the preliminary test are estimated as having a fair chance of passing the final test. Those who score between 19 to 22 may have about 50 percent chance of passing the final test while those scoring 18 or below would have a poor chance of passing (English Language Institute, 2006b).

With regard to the actual test, the speaking and writing sections are scored holistically using guidelines established by ELI-UM.[4] Both sections are scored by trained raters who are continually monitored after their training to ensure proper calibration. For speaking, the scores range from 1 (lowest) to 4 (highest), and the test-taker must receive a score of 2 or above in order to pass the speaking section. For writing, the scores range from A (highest) to D (lowest). The minimum passing score for writing is a C.

The listening and GCVR section items are scored dichotomously based on right or wrong responses. The mean passing scores can vary from year to year because the test forms are different every year, and the difficulty level of the test items in different test forms must be taken into account to determine the passing score. To adjust the difficulty level of the test forms, common items are included to link each form. Then, the English Language Institute uses IRT to determine the results. The cut-off score for the listening and GCVR sections are typically around 60-65 percent as shown in Table 3.4 (English Language Institute, 2006c).

Table 3.4
The ECPE Scoring System
(Adapted from English Language Institute, 2006c)

Section	Honors	Pass	Fail
GCVR (Grammar, Cloze, Vocabulary, and Reading)	Over 90% correct	Above 60-65%	Below 60-65%
Listening	Over 90% correct	Above 60-65%	Below 60-65%
Writing	A	B-C	D
IOC (Speaking)	4	3-2	1

[4] The scoring rubrics for the speaking and writing sections are available on the ELI-UM website: http://www.lsa.umich.edu/UofM/Content/eli/document/ECPE0506InfoBulletin.pdf

The results are sent to the test-takers from the testing centers at which they took the ECPE. ELI-UM does not report the actual ECPE scores. Those who passed the test receive a certificate from the test center.

3.4.3 Coding of the Items

Before statistical analyses were performed in the current study, all the GCVR items were coded to determine what these items were measuring. For the grammar, cloze, and vocabulary items, the coding was based on a model of grammatical knowledge proposed by Purpura (2004), which provides a theoretical definition of grammatical knowledge. As discussed in the literature review section, Purpura suggests that language ability is primarily composed of two elements: grammatical knowledge and pragmatic knowledge. Grammatical knowledge is further divided into two closely related components: grammatical form and grammatical or semantic meaning. Each knowledge component is then defined in terms of six subcomponents at sentential and discourse levels: (1) phonological or graphological form/meaning, (2) lexical form/meaning, (3) morphosyntactic form/meaning, (4) cohesive form/meaning, (5) information management form/meaning, and (6) interaction form/meaning.

Using this model, the items were categorized according to what domain of lexico-grammatical knowledge each item was measuring. The coding was performed by three judges who had been extensively trained in coding grammar items. The coders were given the descriptions of the coding scheme based on Purpura's framework, and asked to classify each item. After the judges individually coded and wrote the grammatical structure for each item, all the codings were recorded on a spreadsheet and compared.

When there was discrepancy in coding, each judge explained the rationale behind coding. For example, if one coder marked an item LF (lexical form) and the other two marked the same item, MF (morphosyntactic form), then the coders discussed the reasons for coding the way they did. After discussion of discrepancies, all the items were ultimately described with a single code. To provide the inter-coder agreement, Fleiss' Kappa (1971) was used as it calculates the extent of agreement among more than two raters. Table 3.5 illustrates the inter-coder agreement for each part of the test.

Table 3.5
Overall Agreement Rates on the GCV Items

Test Form	Grammar	MC Cloze	Vocabulary
Form X	0.83	0.81	0.70
Form Y	0.88	0.77	0.82

The Fleiss' Kappa statistic measuring agreement takes value between 0 and 1, where a value of 1 means complete agreement. The measure ranged from 0.70 to 0.88 with the grammar section being the highest and the vocabulary section being the lowest.

3.4.4 Study Variables in the GCV Section

After numerous discussions on coding, it was determined that the GCV section of the ECPE included items that measured morphosyntactic form (MF), lexical form (LF), lexical meaning (LM), cohesive form (CF), and cohesive meaning (CM). The following section describes the components in Purpura framework, which are used in the current study.

The first component, which is often tested in the grammar items, is morphosyntactic form (MF). As the name of the component suggests, it focuses on a

morphological and/or syntactic form of the language. The features of morphosyntactic form include articles, prepositions, pronouns, inflectional affixes (e.g., -ed), derivational affixes (e.g., un-), simple, compound and complex sentences, mood, and voice. Consider the following example:

I had a hard time _____ for the exam this weekend.

a. studying *
b. to study
c. with study
d. study

* is correct option

In this example, a gerund should be included in the blank. By looking at the sentence and the choices, the test-takers must recognize that the expression "hard time" can only be followed by a gerund complement in this sentence. This item provides the different alternative forms of the same word in order to measure the test-takers' ability to use the appropriate morphosyntactic form, and not lexical meaning.

The second component, Lexical form (LF), allows us "to understand and produce those features of words that encode grammar rather than those that reveal meaning" (Purpura, 2004, p. 92). These include orthography, part of speech (e.g., happ*y*; happ*iness*), morphological irregularity (e.g., go; went), word formation (e.g., nightstand; kickoff), countability (e.g., children; people) / gender (e.g., actress) restrictions, formulaic expressions (e.g., You're welcome) and co-occurrence restrictions (e.g., attract *to*, *in* spite *of*). A co-occurrence restriction occurs when a verb or a transitive adjective is followed by a particular preposition (e.g., depend *on* X; yield *to* X) or a given noun phrase is preceded by a particular preposition (e.g., *in* my opinion) (Celce-Murcia & Larsen-Freeman, 1999). The following is an example of LF, which demonstrates a co-occurrence restriction:

Mary gets along ____ her roommates well.

a. with *
b. of
c. for
d. to

* is correct option

The correct option is *with*. In this example, the phrase *get along* is followed by the preposition *with*. This is considered the grammatical dimension of lexis, representing a co-occurrence restriction with prepositions (Purpura, 2004).

The third component, lexical meaning (LM) is closely associated with LF. A difference between the two is that LF focuses on the grammatical structure of a word, whereas the LM emphasizes the literal meaning of a word. Consider the following example:

There's a serious ____ between the two university football teams.

a. competition *
b. bile
c. temper
d. exasperation

* is correct option

All four choices for the blank are nouns; thus, this item is not measuring the form of the word. Instead, it is examining whether the test-takers understand the meaning of the word in context. The word, *competition*, carries the meaning of rivalry and is the correct choice in this example.

A forth component measured in the GCV section is cohesive form (CF). According to Purpura (2004), "knowledge of cohesive form enables us to use the phonological, lexical and morphosyntactic features of the language in order to interpret and express cohesion on both the sentential and the discourse levels" (p. 95). This

includes cohesive devices such as pronoun referents, and ellipses (e.g., so do I; I do too). Consider the following example:

The woman _____ is wearing glasses is my sister.

a. who *
b. whose
c. whom
d. which

* is correct option

All four options are relative pronouns; however, there are some distinctive features the test-takers need to know in order to get this item correct. First, all relative pronouns have different syntactic properties. The relative *who* and *which* are subjective cases while *whom* and *whose* are objective cases. Moreover, relative *whose* is a possessive determiner, which typically refers to a human head noun. Second, these relative pronouns have different semantic properties. The relative pronouns, *who*, *whose*, and *whom* have [+human] while *which* has [-human].

By considering the distinctive features of each pronoun, the test-taker may arrive at one correct answer. The clause, *the woman _____ is wearing glasses* is a subject NP, thus it only takes a subjective case: *who* or *which*. Now consider the distinguishing characteristic of *who* and *which*. The pronoun, *which*, would be incorrect in this case because it is not semantically linked to its referent, *the woman*. The pronoun, *who*, on the other hand, has the same semantic property [+human] as its referent, *the woman*. Hence, the correct answer is *who*. In order to answer this item correctly, the test-taker needs to know *the woman* and the relative *who* are co-referential, which is part of the definition of cohesive form.

The last component used in this study is cohesive meaning (CM). Purpura (2004) states that CM and CF are closely associated through cohesive devices that create

connections between cohesive forms and their referential meanings within the linguistic environment or the surrounding co-text. CM can be conveyed through substitution (e.g., *I guess so*), logical connectors (e.g., *therefore*; *however*) and lexical connection in the form of synonymy and repetition. Consider the following example:

> Susan is not one of my close friends, _____ I don't think I'll invite her to my birthday party.
>
> a. so*
> b. but
> c. or
> d. yet
>
> * is correct option

All four choices are conjunctions, so this test item is not testing form. In order to answer this item correctly, the test-taker needs to understand both the first and second part of the sentence and choose the appropriate connector. In other words, the test-taker needs to understand the cohesion between the two clauses. The first part of the sentence says Susan is not a good friend. Then, the second part of the sentence says, "I'm not going to invite her to my party." These two clauses are connected in the pattern of cause and result, and *so* is the only marker which functions this way. Hence, this is the only correct option.

After the coding was completed, the items were categorized in terms of the five components listed above: lexical form (LF), lexical meaning (LM), morphosyntactic form (MF), cohesive form (CF), and cohesive meaning (CM). Table 3.6 presents the taxonomy of the GCV items in the two test forms. Grammar items are indicated with a letter G in front of the item number. For example, grammar item 13 is indicated as G13. Similarly, C refers to a cloze item, and V to vocabulary item. For both Form X and Y, six items were coded as LF, twenty items as MF, eight items as CF, forty-three items as LM, and three items as CM.

Table 3.6
Taxonomy of the GCV Items in the Two Test Forms

Components (GCV items)	Number of Items	Items
Form X	Total: 80	
Lexical Form (LF)	6	G2, G9, G12, G15, C7, C13
Morphosyntactic Form (MF)	20	G3, G4, G5, G7, G8, G10, G13, G14, G17, G18, G19, G20, G21, G23, G24, G27, G28, G29, G30, C10
Cohesive Form (CF)	8	G6, G11, G16, G22, G25, G26, C5, C14
Lexical Meaning (LM)	43	G1, C1, C2, C3, C6, C8, C9, C15, C16, C17, C18, C19, C20, V1, V2, V3, V4, V5, V6, V7, V8, V9, V10, V11, V12, V13, V14, V15, V16, V17, V18, V19, V20, V21, V22, V23, V24, V25, V26, V27, V28, V29, V30
Cohesive Meaning (CM)	3	C4, C11, C12
Form Y	Total: 80	
Lexical Form (LF)	6	G3, G7, G23, G30, V4, V14
Morphosyntactic Form (MF)	20	G1, G2, G4, G8, G9, G10, G11, G12, G13, G15, G18, G19, G20, G21, G24, G25, G26, G27, G29, C20
Cohesive Form (CF)	8	G5, G6, G14, G17, G22, C1, C7, C13
Lexical Meaning (LM)	43	G16, G28, C2, C3, C4, C5, C6, C8, C11, C12, C15, C16, C17, C18, C19, V1, V2, V3, V5, V6, V7, V8, V9, V10, V11, V12, V13, V15, V16, V17, V18, V19, V20, V21, V22, V23, V24, V25, V26, V27, V28, V29, V30
Cohesive Meaning (CM)	3	C9, C10, C14

3.4.5 Content Analysis of the GCV Section

After the coding was determined, the coders were asked to further determine the grammatical structures of each item. Consider the following example:

I _____ up all night last night studying for a math exam.

a. stayed *
b. stay
c. am staying
d. will stay

* is correct option

This item measures the test-taker's ability to understand the use of past tense. Hence, a coder should write down, "verb tense, past" to specify the assessed grammatical feature of this item. It was important to record the grammatical feature along with the coding for each item, because Form X and Form Y could be similar in terms of the number of coded items in each category of lexico-grammatical knowledge, but different in terms of grammatical features tested. Table 3.7 shows the summary of grammatical features measured in the GCV section of Form X and Y.

There were some grammatical features tested only in one form and not in the other. For example, there were items measuring pronouns and questions (e.g., why-, y/n, tags) in Form X while there were none in Form Y. Furthermore, conditionals, emphasis, and formulaic expressions were only tested in Form Y. Overall, many grammatical features were tested in both forms in relatively equal amounts.

Table 3.7
Grammatical Features Tested in the GCV Section of Form X and Form Y

Number of Items		Tested Grammatical Features
Form X	Form Y	
11	12	Adjectives and adjective phrases
8	7	Adverbs and adverbials
3	3	Complements (DO + IO) and complementation
0	2	Conditionals
0	2	Focus and emphasis (e.g., cleft)
0	3	Formulaic expressions
5	3	Logical connectors and conjunctions
2	0	Modals and phrasal modals (e.g., have to)
16	10	Nouns and noun phrases
0	2	Passive voice
2	3	Phrasal verbs
4	2	Prepositions and prepositional phrases
2	0	Pronouns and reference
1	0	Questions (e.g., wh- y/n, tags) and answers
2	1	Relative clauses
22	27	Tense and aspect; other verb forms (e.g., past part)
2	3	Word order
Total 80	80	

3.4.6 Study Variables of the Reading (R) Items

The reading items were coded based on the four components discussed in the literature review section: (1) understanding main idea (MAIN), (2) reading for detail information (DET), (3) understanding vocabulary in context (VOC), and (4) reading for inferential information (INF). The same judges who coded the grammar, cloze, and vocabulary items coded the reading items. After the judges individually coded the items, all the codings were again recorded on a spreadsheet and compared. When there was a discrepancy in the coding, each coder again explained the rationale behind the coding. For example, if one coder marked MAIN and the other two marked INF, then the coders discussed the reasons for coding the way they did. After discussion of discrepancies, all the items were ultimately described with a single code. To provide the inter-coder

agreement, Fleiss' Kappa (1971) was used as it calculates the agreement of more than two raters. Table 3.8 illustrates the inter-coder agreement using Fleiss' Kappa for the reading part of the test.

Table 3.8
Overall Agreement Rates on Reading Items

Test Form	Reading
Form X	0.88
Form Y	0.94

The Fleiss' Kappa statistic measuring agreement takes value between 0 and 1, where a value of 1 means complete agreement. The agreement was 0.88 for Form X and 0.94 for Form Y. The results suggest that both forms had high agreement rate, especially for Form Y.

Table 3.9 presents the taxonomy of the reading items in Form X and Form Y. Form X contains two main idea questions, eleven detail questions, one vocabulary question, and six inference questions. Form Y also contains two main idea questions, eleven detail questions, one vocabulary question, and six inference questions.

Table 3.9
Taxonomy of the Reading Items in the Two Test Forms

Components (Reading Items)	Number of Items	Items
Form X	Total: 20	
Main Idea (MAIN)	2	R5, R6
Detail (DET)	11	R1, R2, R3, R10, R11, R13, R15, R16, R17, R18, R19
Vocabulary in Context (VOC)	1	R7
Inference (INF)	6	R4, R8, R9, R12, R14, R20
Form Y	Total: 20	
Main Idea (MAIN)	2	R6, R11
Detail (DET)	11	R1, R2, R3, R5, R7, R8, R15, R16, R17, R18, R19
Vocabulary in Context (VOC)	1	R12
Inference (INF)	6	R4, R9, R10, R13, R14, R20

3.5 Analyses

This section discusses the computer equipment and software used in the analysis. It then describes the statistical procedures used to analyze the data in this study.

3.5.1 Computer Equipment and Software

First, Microsoft EXCEL for the PC Version 10.0 was used to input the data. These data then exported to other statistical programs. To compute descriptive statistics and to perform reliability analyses and exploratory factor analyses, SPSS Version 12.0.0 for the PC (SPSS Inc., 2003) was utilized. Finally, EQS 6.1 (Bentler, 2006) was used to perform confirmatory factor analyses and multi-group structural equation modeling.

3.5.2 Descriptive Statistics and Assumption Checking

To examine the normality assumption, descriptive statistics (e.g., the mean, median, and standard deviation) for each of the test items were calculated. The kurtosis and skewness of each variable were also calculated in order to examine the item distribution. Univariate and multivariate outliers were identified, and if there were any extreme outliers were, they were deleted and reported. Assumptions regarding univariate, multivariate normality, and linearity were examined, because these assumptions are required for implementation of the maximum likelihood parameter estimation method utilized in confirmatory factor analysis (Park, 2007).

3.5.3 Reliability Analysis

To examine consistency of measurement and the internal structure of the test, the internal consistency reliability estimates (i.e., the coefficient alpha) were calculated for each section and for the GCVR section. The item-total correlations for each item as well as the overall estimate of the scale reliability were also investigated in order to examine the homogeneity of the items within each section. For the cloze section, Guttman's split-half procedure was used to perform reliability analysis because the cloze items may violate the assumption of independence (Bachman, 1990).

3.5.4 Structural Equation Modeling

Subsequent to the reliability analysis, the structural equation modeling (SEM) analytic procedures were performed in order to examine the relationship between observed and latent variables, and among latent variables. SEM is an integration of

multiple regression, path analysis, multitrait-multimethod analysis, and factor analysis. The incorporation of these statistical analyses allows researchers to separately and simultaneously examine the links between the latent factors and their observed measures as well as the links among the latent factors (Bollen & Long, 1993; Byrne, 2006; Kunnan, 1998).

There are three basic approaches to SEM (Jöreskog, 1993): (1) strictly confirmatory approach; (2) alternative models approach; and (3) model-generating approach. In the strictly confirmatory approach, a researcher postulates a single model and examines the extent to which it is consistent with the data. Based on how well the model fits the data, the researcher either accepts or rejects the model. A limitation to this approach is that the scope of model testing is so narrow that it leaves little flexibility to consider other unexamined models which may fit the data as well or better than the hypothesized model (Kline, 1998). In the alternative models approach, a researcher tests two or more hypothesized models and determines which model best fits the data (Kunnan, 1998). This approach is often used when the researcher wants to compare competing theories or examine contradictory research findings found in the literature. The most common approach used in SEM is the model-generating approach (Kline, 1998). This approach is used when an initial model does not fit the data and the model is revised to improve its parsimony. The current study used the model-generating approach to test a hypothesized model.

To illustrate a hypothesized model and show the results of the analysis, researchers often use a diagram. There is a general graphing convention that has been adopted over the years, but there are some different types of symbols used depending on

statistical software applications. This study used the Bentler-Weeks (1980) representation system. In this system, all variables in a model can be grouped as either dependent or independent variables. Variables that have unidirectional arrows pointing to them are called dependent variables, while variables that have no directional arrows pointing at them are called independent variables. To illustrate the relationship between these variables, there are mainly four symbols used in model diagramming (see Table 3.10).

Table 3.10
Symbols Used in Bentler-Weeks (1980) Representation System
(Adopted from Bentler, 2006)

Symbol	Name	Meaning
V	Variable	Measured variable
F	Factor	Latent variable
E	Error	Residual of measured variable
D	Disturbance	Residual of latent variable

All measured or observed variables are labeled as V's and denoted by rectangles. All latent variables are labeled as F's to refer to factors and denoted by ovals. Errors related to the measurement of each observed variable are denoted as E's. Errors related to the prediction of latent variables are denoted as D's to refer to disturbances.

Steps in Structural Equation Modeling

There are mainly five steps in an SEM application (Bollen & Long, 1993): (1) model specification, (2) model identification, (3) model estimation, (4) testing model fit, and (5) model respecification. Based on the substantive theory of grammatical knowledge and reading ability described in Chapter 2, structural equation models are specified. The following section discusses and applies each step of the procedures in the current study.

Step 1: Model Specification

Based on substantive theory and empirical research, a researcher formulates a hypothesis, examining a series of relationships among observed and latent variables. SEM generally involves two types of models: a measurement model and a structural model. The measurement model specifies the extent to which the latent variables are measured in terms of the observed variables. The structural model specifies the relationships among latent variables. The model which comprises both a measurement model and structural model is called full latent variable model (Byrne, 2006). In the present study, both types of models are generated to investigate the underlying constructs of lexico-grammatical knowledge and reading ability measured by ECPE. There were four hypothesized models in order to answer the research questions outlined in Chapter 1. The models in the current study were presented following the discussion on the steps in structural equation modeling.

Step 2: Model Identification

After specifying each respective model, the identification of the model needs to be checked. According to Byrne (2006), "the issue of identification focuses on whether there is a unique set of parameters consistent with the data" (p. 31). As Schumacker and Lomax (1996) state, "model identification ... depends on the specification of parameters as fixed, free, or constrained. Once the model is specified and the parameter specifications are indicated, the parameters are combined to form one and only one $\sum$ (model-implied variance-covariance matrix)" (p. 100). A problem with identification may

still exist, which depends in part on the amount of information in the covariance matrix essential for the unique estimation of the parameters of the model (Kunnan, 1998).

There are three levels of model identification (Byrne, 2006; Kline, 1998). First, a model is said to be *under-identified* if one or more parameters may not be uniquely determined from the covariance matrix. Second, a model is referred to as *just-identified* if all of the parameters are uniquely estimated from the covariance matrix. Third, a model is called *over-identified* when there is more than one way to determine the parameter(s). Theoretically, when the model is not identified (*under-identified*), only some of the parameters are estimated and the model is not trustworthy. In a study utilizing SEM, one of the aims is to specify an *over-identified* model, in which results positive degrees of freedom that allow for rejection of the model (Byrne, 2006). In the current study, all hypothesized models were over-identified; hence, it was possible to proceed with model estimation.

Step 3: Model Estimation

The purpose of model estimation is to obtain estimates for each of the parameters specified in the model (Schumacker & Lomax, 1996). More specifically, SEM estimates the unknown or free parameters in the model from the given sample data. Although there are many types of estimation procedures, the current study used the maximum likelihood (ML) method, with the assumption of a multivariate normal distribution of the data. In order to examine the validity of the assumption of multivariate normality, the EQS program computes Mardia's (1970, 1974) normalized multivariate kurtosis coefficient. Bentler (2006) suggested that the values "should be roughly in the +3 to -3 range, though

somewhat larger values are probably not too worrisome" (p. 129). As part of the analysis in the present study, this kurtosis statistic was examined as a test of the appropriateness of the normality assumption.

The data for the current study did not meet the assumptions of the multivariate normality; hence, the ML robust estimation method was used. This estimation method uses the Satorra-Bentler scaled test statistic (1994), which corrects the test statistic and standard errors of the normal theory estimator (Bentler, 2006).

Step 4: Testing Model Fit

The purpose of model fit is to determine the degree to which the model fits the sample data. Bollen and Long (1993) recommended that multiple measures be reported as each fit index has its limitations. Hence, this study used multiple fit indices to ensure that the generated models in the study fit the data statistically, as well as substantively, for all estimated parameters. The types of goodness of fit criteria used in this study were: χ^2, the parsimony ratio (χ^2/df), Comparative Fit Index (CFI), Bentler-Bonnet Non-Normed Fit Index (NNFI), Root Mean Square Error of Approximation (RMSEA).[5] The statistical criteria and fit indices recommended by Hu and Bentler (1999) were used in the current study (see Table 3.11).

[5] See Schumacker and Lomax (1996, p. 121, Table 7.1) for a table of goodness of fit criteria and acceptable fit interpretation.

Table 3.11
Statistical Criteria and Fit Indices for Model Fit
(Adapted from Chang, 2004)

Fit Index	Criterion
χ^2 Goodness-of-Fit Test	Low and non-significant values
Parsimony Ratio (χ^2/df)	Ratio of 3 or less
Bentler Comperative Fit Index (CFI)	0.95 or above
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.95 or above
Root Mean Square Error of Approximation (RMSEA)	0 .05 or less

The χ^2 statistic examines the specified model against the unconstrained or null model, and then measures the difference between the observed sample covariance matrix and the model-implied covariance matrix (Jöreskog and Sörbom, 1989). A statistically significant χ^2 value relative to the degrees of freedom indicates that the difference between the observed and estimated matrices is due to sampling variation. On the other hand, a statistically nonsignificant χ^2 value indicates that there is no difference between the observed and estimated matrices. This implies that the data fit the model, but it is uncertain that other models might not have similar model fits (Schumacker & Lomax, 1996). Thus, a researcher would be interested in obtaining a nonsignificant χ^2 value with associated degrees of freedom.

Another known feature of χ^2 statistic is that it is sensitive to sample size (Byrne, 2006). When sample size is large, even a small discrepancy between a model and its sample data can result in the rejection of a well-fitted model. Given the large size of the sample in this study, the sample size affected the χ^2 statistic. Therefore, other fit indices were also considered to determine the model fit.

The Comparative Fit Index (CFI) (Bentler, 1990) is a revised version of the Normed Fit Index (NFI) which compares a hypothesized model with a null model and

presents a measure of complete covariance in the data. Unlike NFI, CFI is not dependent on sample size; thus, Bentler (1990) recommends that the CFI should be the index of choice. Based on a 0 to 1.0 scale, a CFI value greater than 0.95 indicates acceptable indices of model fit (Hu & Bentler, 1999).

Bentler-Bonnet Non-normed Fit Index (NNFI) has similar characteristics of CFI. The difference is that the NNFI penalizes model complexity and awards model parsimony, while the CFI does not take the model parsimony into account (McDonald & Marsh, 1990). A NNFI value greater than 0.95 indicates acceptable indices of model fit (Bentler, 1990). Although the scale is based on 0 to 1.0, the NNFI can exceed 1.0 due to sampling fluctuations. Hence, this index should be used with other indices and not as a sole determiner of a model fit.

The Root Mean Square Error of Approximation (RMSEA) takes into account the error of approximation in the population (Byrne, 2006). This is a commonly used measure of fit, partly because it does not require comparison with a null model (Garson, n.d.). By convention, the model is considered good fit if the RMSEA is less than or equal to 0.05, and it is considered adequate fit if the RMSEA is less than or equal to 0.08. Hu and Bentler (1999), however, suggested that RMSEA with a value of 0.06 is the cutoff for a good model fit. For the purpose of the present study, an RMSEA value less than 0.05 was considered a good model fit.

Step 5: Model Respecification

When all the goodness of fit indices are in an acceptable range, the model is considered well-fit. When the fit indices are not in the acceptable range, the model is

considered misfitting. In that case, the model must be respecified. In order to detect specification errors, EQS offers several statistical procedures. The current study used the Lagrange Multiplier (LM) test and the Wald Test.

The LM test shows the degree to which the model fit would increase if the fixed parameters were to be freely estimated. In other words, the LM test examines the effect of adding free parameters to a restricted model (i.e., reducing restriction on the model) (Bentler, 2006). The Wald test, on the other hand, shows the degree to which the model fit would increase if the free parameters were to be fixed to zero or dropped from the model (Kline, 1998). In other words, the Wald test examines "the effect of dropping free parameters from a more complete model (i.e., adding restrictions to the model)" (Bentler, 2006, p. 159).

Although these procedures are helpful in identifying specification errors and improving a model fit, they should be used only when there is a substantively meaningful theory to support the model respecification. In the present study, SEM was performed as a *model-generating* (Jöreskog, 1993) procedure, in which an initial model was identified and tested. When the model did not fit the data, the model was modified and tested until a model with a theoretical rationale and an acceptable statistical fit was generated.

3.6 Models in the Current Study

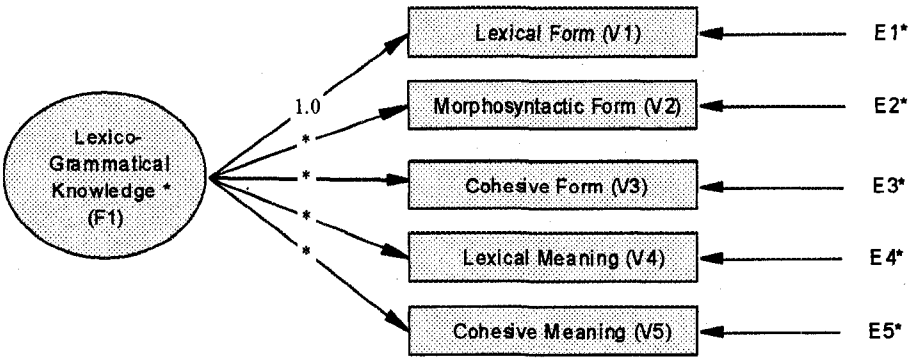
In the present study, there were four hypothesized models postulated to investigate the underlying constructs of lexico-grammatical knowledge and reading ability measured by ECPE. There were primarily two different types of analyses involved in examining the models: single-group and multi-group analyses. As the name suggests,

single-group analyses only involve single group of population whereas multi-group analyses involve multiple groups of population (e.g., low-ability group vs. high-ability group or Spanish speaking group vs. English speaking group). In the multi-group analyses, the main goal is to examine the degree to which the models are variant across different groups. In order to perform multi-group analyses, it is necessary to establish a separate baseline model for each group. The following section describes each hypothesized model used in this study.

3.6.1 Single-Group Analyses

As seen in Figure 3.2, the first hypothesized model under investigation was the construct of lexico-grammatical knowledge measured by grammar, cloze, and vocabulary (GCV) section of ECPE. Based on substantive theory and the operationalized definition of the GCV section, lexico-grammatical knowledge (F1) was measured by five observed variables called lexical form (V1), morphosyntactic form (V2), and cohesive form (V3), lexical meaning (V4), and cohesive meaning (V5). This model was a measurement model as it examined the relationship between a latent variable (i.e., lexico-grammatical knowledge) and observed variables. It is important to note that Purpura (2004) never postulated that lexico-grammatical knowledge was or was not a measurement model. This study initially treated Purpura's model as a measurement model. This model was analyzed for both Form X and Y.

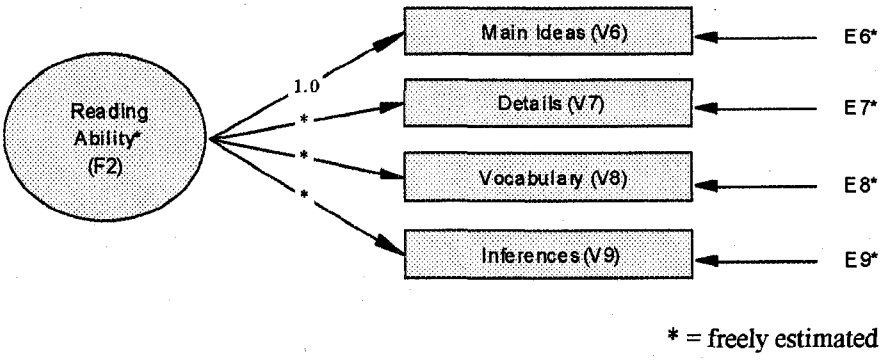
Figure 3.2
Hypothesized Model for the GCV Section



* = freely estimated

The second initially hypothesized model under investigation was the construct of reading ability measured by reading section of ECPE. This model was also a measurement model as it examined the relationship between a latent variable (i.e., reading ability) and observed variables (i.e., main ideas, details, vocabulary, and inferences). As illustrated in Figure 3.3, this model asserted that the variance in reading ability was explained by the test items which measure: the ability to identify main ideas (V6), the ability to identify details (V7), the meaning of vocabulary in context (V8), and the ability to identify inferences (V9). This model was analyzed for both Form X and Form Y.

Figure 3.3
Hypothesized Model for the Reading Section

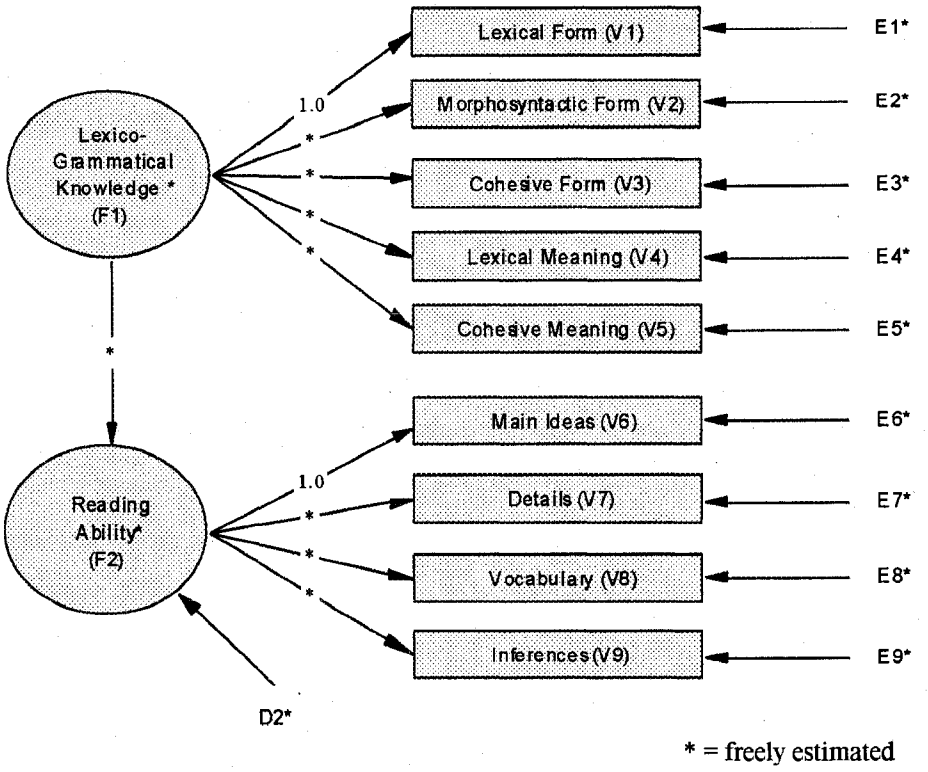


* = freely estimated

The third initially hypothesized model under investigation included the constructs of lexico-grammatical knowledge and reading ability measured by the GCVR section of the ECPE. This model was generated by specifying a relationship between the two latent variables: lexico-grammatical knowledge and reading ability. This model postulated that reading ability is regressed from lexico-grammatical knowledge. This idea has been taken from Purpura's (1999) study, which examined the relationship between lexico-grammatical ability and reading ability on First Certificate of English (FCE) Anchor test developed by University of Cambridge Local Examinations of Syndicates (UCLES). Purpura's (1999) study showed that reading and lexico-grammatical ability not only correlated with one another, but lexico-grammatical ability had a strong impact on reading ability. Grabe (2005) also supports the idea that grammar resources are central to fluent reading ability. Similarly, Bernhardt (1999) reviewed the studies (e.g., Bernhardt & Kamil, 1995; Bossers, 1991; Brisbois, 1995; Carrell, 1991) which investigated the contribution of L1 reading and L2 grammar to L2 reading, and concluded that L2 reading is primarily dependent on grammatical ability in the second language. These empirical studies indicate that reading ability requires a certain threshold of lexico-grammatical knowledge in order to understand syntactic structure as well as literal and intended

meaning. Based on this empirical evidence, the current study hypothesized that lexico-grammatical knowledge was a critical linguistic resource for reading ability. The diagrammatic representation of this model is shown in Figure 3.4. This structural model was analyzed for both Form X and Form Y.

Figure 3.4
Hypothesized Model for the GCVR Section



3.6.2 Multi-Group Analyses: Comparing Forms

Following the establishment of the baseline model, simultaneous multi-group analyses were performed. The main goal in multi-group analyses was to examine the degree to which the models were equivalent across groups. According to Jöreskog (1971), the analysis should start with a test of the equality of covariance structures across groups. To do so, the null hypothesis must be identified: H_o: $\Sigma_1 = \Sigma_2 = \Sigma_3 = \ldots \Sigma_g$ where Σ is the

population variance-covariance matrix and g is the number of groups. The null hypothesis states that all groups are identical. If the null hypothesis is rejected, there is no equivalence among groups and that there is a need for more restrictive hypothesis to identify the source of noninvariance (Byrne, 2006). If the null hypothesis cannot be rejected, groups are considered to be equivalent in covariance structures. In that case, there is no need for further tests for invariance.

The current study examined the extent to which the GCVR section measured the same underlying trait structures across Form X and Form Y. The initial step in testing for invariance was to examine the configural invariance. This means that the same factor structure must hold across forms. To test for configural invariance, the baseline model for Form X and Form Y were estimated simultaneously without any equality constraints are imposed on the parameters. The model tested here is a multi-group representation of the baseline models.

The next step in testing for invariance is to determine equality with respect to the measurement model. In this test, the invariance of factor loadings is of interest. To test for multi-group invariance across forms, between-form equality constraints were imposed on the parameters to be tested for invariance. The null hypothesis tested for between-form invariance of the parameters of interest. The null hypothesis tested if parameter *z* in Form X were equal to parameter *z* in Form Y. If $p<0.05$, the null hypothesis was rejected, indicating that parameters in question were variant across forms. On the other hand, if $p\geq0.05$, the null hypothesis could not be rejected, indicating that the parameters were not considered to be variant across forms. With EQS application, the Lagrange Multiplier Test serves as a means of examining the univariate and multivariate tenability of these

cross-group (form) equality constraints (Bentler, 2006; Purpura, 1999). Tests for the measurement error variances-covariances are not generally examined as it is considered excessively stringent (Byrne, 2006). Hence, the current study focused on the invariance of factor loadings, and the errors for the observed variables were not constrained to be equal.

The last step in testing for invariance is to address the equality with respect to the structural model. In this test, the factor variances are constrained to be equal across forms. In order to simultaneously estimate the models across forms, a series of cross-group equality constraints on all parameters of interest was imposed (Purpura, 1999). These parameters included the cross-form invariance of the factor loadings (F→V) in the measurement models, the cross-form invariance of the regression paths (F→V) in the structural models, and the cross-form invariance of the factor variances (Purpura, 1999). Once the equality constraints were imposed on the parameters, the models were estimated to see how well they fit the data.

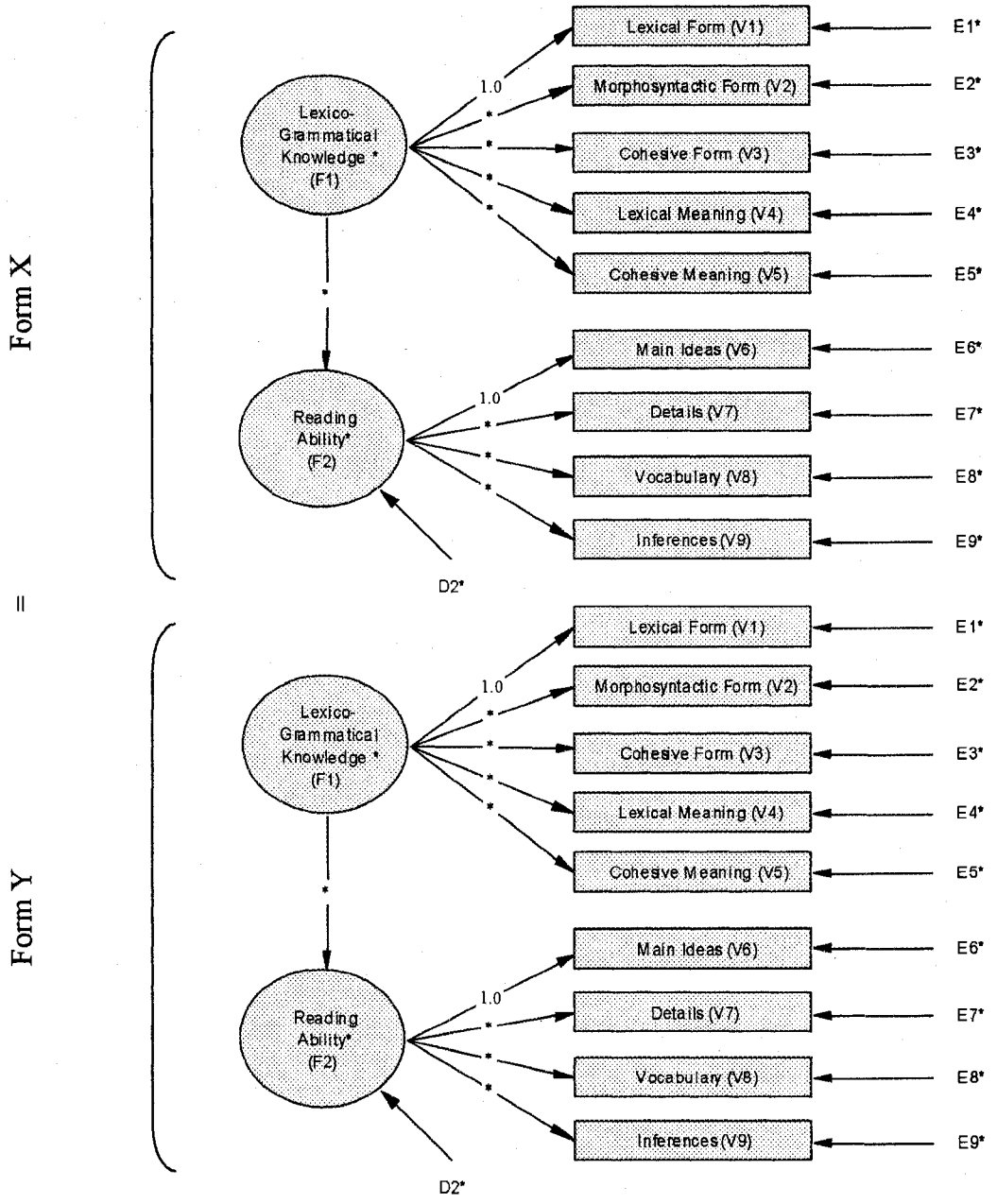
In sum, the multi-group analysis investigates the extent to which the model fits across two test forms. This was used to provide evidence of comparability of the two test forms of the GCVR section. The diagrammatic representation of this model is shown in Figure 3.5.

3.7 Summary

This chapter outlined the research methodology that was used in the present study. It described the participants, the measurement instrument, and the methods used to code

the items. The statistical procedures used to analyze the data were then explained. In the next chapter, the analyses of the present study are discussed.

Figure 3.5
Hypothesized Model for the GCVR Section across Two Test Forms



* = freely estimated

Chapter IV

PRELIMINARY ANALYSES

Prior to examining the underlying trait structures of the GCVR section of Form X and Form Y, a number of preliminary analyses were performed. In this chapter, the grammar, cloze, vocabulary (GCV) section is separately analyzed from the reading section. This is due to the results of the pilot study (Saito, 2003) as well as the findings in the ECPE annual report (English Language Institute, 2006b, 2006c). The previous study results showed that the reading section clearly measured a separate construct from the ability measured in the GCV section. Hence, the GCV section was examined independently from the reading section.

Both the GCV and reading sections are analyzed in two steps. First, the section-level distribution of each section, and the distribution of the individual items based on theoretical coding are examined. Second, the reliability estimates for each section and the variables used for coding are examined.

The results in this chapter are valuable in accurately modeling the underlying trait structure of the lexico-grammatical knowledge and reading ability measured in the GCVR section of the ECPE.

4.1 GCV Section

In this section, the GCV sections of Form X and Form Y were analyzed based on the responses of 33,662 and 32,473 test-takers respectively. Descriptive statistics and reliability analyses are presented.

4.1.1 Section-Level Distributions of the GCV Section

The distributions of each section as a whole were examined. Table 4.1 presents the summary distributions of the GCV section for Form X and Y.

The mean for the grammar section was 22.32 for Form X and 22.05 for Form Y based on 30 items. The standard deviation was 3.91 and 4.01 respectively. The values of the means and the standard deviations indicate that the difficulty of the items and the dispersion of the test data were very similar across forms. Skewness and kurtosis were within the acceptable limits.

The mean for the MC cloze section was 14.90 for Form X and 12.51 for Form Y based on 20 items. This indicates that the MC cloze items on Form X were a little easier than on Form Y. The standard deviation was 2.50 for Form X and 3.30 for Form Y, which indicates that there was a wider variation in the test results for Form Y. Skewness and kurtosis were within the acceptable limits.

The mean for the vocabulary section was 17.68 for Form X and 17.92 for Form Y based on 30 items. The standard deviation was 4.33 and 4.46 respectively. The values of the means and the standard deviations indicate that the difficulty of the items and the dispersion of the test data were very similar across forms. Skewness and kurtosis were within the acceptable limits.

The mean for the GCV section was 54.90 for Form X and 52.48 for Form Y based on 80 items. The differences in the means for the GCV section were mostly contributed by the difference in the means for the MC cloze section as the mean differences in the grammar and vocabulary sections were marginal. The standard deviation was 8.49 for Form X and 9.75 for Form Y, which indicates that there was a wider variation in the test results for Form Y. Again, the differences in the standard deviations for the GCV section were mostly contributed by the differences in the standard deviations for the MC cloze section. Skewness and kurtosis were within the acceptable limits.

Table 4.1
Distributions of the GCV Section

	# of	Mean		Stdv		Skewness		Kurtosis	
Section	Items	X	Y	X	Y	X	Y	X	Y
Gram (G)	30	22.32	22.05	3.91	4.01	-0.52	-0.50	0.21	0.10
Cloze (C)	20	14.90	12.51	2.50	3.30	-0.58	-0.19	0.48	-0.36
Voc. (V)	30	17.68	17.92	4.33	4.46	0.04	-0.04	-0.21	-0.32
GCV Total	80	54.90	52.48	8.49	9.75	-0.18	-0.13	0.33	-0.11

In order to compare means of the GCV sections across Form X and Form Y, t-test was calculated using SPSS 12.0. Table 4.2 presents the summary of the t-test. The results show that the differences in means for all sections were significantly different. In other words, Form X and Form Y were significantly different[6]. However, it should be noted that even a tiny numerical difference in the means can be significantly different considering the large sample size of this data. Furthermore, there is not much meaningful discrepancy in the mean difference; hence, it would be premature to conclude that Form X and Form Y have different underlying structures based on the results of the t-test.

[6] Based on the information on the participants described in Chapter 3, it is assumed that the populations of test-takers are identical across the two tests. Therefore, the differences in means are interpreted as a difference in the tests.

Table 4.2
T-test Results for the GCV Section

Section	t-test for equality means				
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Gram (G)	8.63	66133	0.00	0.27	0.03
Cloze (C)	105.17	66133	0.00	2.39	0.02
Voc. (V)	-6.95	66133	0.00	-0.24	0.03
GCV Total	34.03	66133	0.00	2.42	0.07

4.1.2 Distributions of the GCV Items Based on Theoretical Coding

Following the section-level distribution analysis, the test items were grouped together into five conceptual categories: lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning. These categories were based on the theoretical coding (Purpura, 2004) described in Chapter 2. Table 4.3 presents the summary distribution of these five category variables for Form X and Form Y.

The mean for lexical form was 4.82 for Form X and 3.41 for Form Y based on six items. This indicates that the lexical form items measured in Form X were easier than the lexical form items measured in Form Y. The standard deviation was 1.05 for Form X and 1.31 for Form Y, which indicates there was not much difference in the dispersion of the data. Skewness and kurtosis were within the acceptable limits for the lexical form items.

The mean for morphosyntactic form was 15.29 on both Form X and Form Y based on twenty items, which indicates that the difficulty level of the morphosyntactic form items was the same in both forms. The standard deviation was 2.82 for Form X and 2.87 for Form Y, which suggests there was not much difference in dispersion of the data. Skewness and kurtosis were within the acceptable limits for the morphosyntactic form items.

The mean for cohesive form was 5.17 for Form X and 6.00 for Form Y based on eight items. This indicates that cohesive form items measured in Form Y were slightly easier than the cohesive form items measured in Form X. The standard deviation was 1.40 for Form X and 1.28 for Form Y, which indicates the difference in the dispersion was minimal. Skewness and kurtosis were within the acceptable limits for the cohesive form items.

The mean for lexical meaning was 27.42 for Form X and 26.07 for Form Y based on 43 items. This indicates that the lexical meaning items measured in Form X were slightly easier than the lexical meaning items measured in Form Y. The standard deviation was 5.22 for Form X and 5.95 for Form Y, which indicates there was not much difference in the dispersion of the data. Skewness and kurtosis were within the acceptable limits for the lexical meaning items.

The mean for cohesive meaning was 2.20 for Form X and 1.71 for Form Y based on three items. This indicates that the cohesive meaning items measured in Form X were slightly easier than the cohesive meaning items measured in Form Y. The standard deviation was 0.80 for Form X and 0.90 for Form Y, which indicates there was not much difference in the dispersion of the data. Skewness and kurtosis were within the acceptable limits for the cohesive meaning items.

Table 4.3
Distributions of the GCV Items Based on Theoretical Coding

Variables	# of items	Mean		Std Dev		Skewness		Kurtosis	
		X	Y	X	Y	X	Y	X	Y
Lexical F.	6	4.82	3.41	1.05	1.31	-0.78	-0.13	0.37	-0.40
Morph F.	20	15.29	15.29	2.82	2.87	-0.66	-0.61	0.34	0.18
Cohesive F.	8	5.17	6.00	1.40	1.28	-0.27	-0.05	-0.13	-0.01
Lexical M.	43	27.42	26.07	5.22	5.95	-0.01	-0.04	0.01	-0.24
Cohesive M.	3	2.20	1.71	0.80	0.90	-0.67	-0.21	-0.32	-0.75

In order to compare means of the variables across Form X and Form Y, t-test was calculated. Table 4.4 presents the summary of the t-test. The results show that the differences in means for all variables, except morphosyntactic form, were significant. The means for morphosyntactic form were identical (mean = 15.29) across Form X and Form Y. The t-test confirms that the means of morphosyntactic form variable on Form X and Form Y were the same.

As for the remaining variables, the t-test results were statistically significant, indicating that the variable means on Form X and Form Y were different. Under the assumption that the test-taking populations for the two forms are identical, this would mean the forms are at least somewhat different. It should again be noted; however, that even numerically small differences in the means may be significantly different considering the large sample size of this data. The differences in means were very large for the lexical form variable, the difference in the mean on the two forms is more than a full standard deviation in each the form-specific category scores. For the other variables, the differences are not nearly so large, relative to the standard errors of the underlying variables. While there is strong evidence that something is different about the lexical form category, it would again be premature to conclude that Form X and Form Y have different underlying structures based on the results of the t-test.

Table 4.4
T-test Results for the GCV Items Based on Theoretical Coding

Variables	t-test for equality means				
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Lexical F.	152.73	66133	0.00	1.41	0.01
Morph F.	-0.03	66133	0.97	0.00	0.02
Cohesive F.	-79.58	66133	0.00	-0.83	0.01
Lexical M.	30.88	66133	0.00	1.34	0.04
Cohesive M.	7.85	66133	0.00	0.50	0.01

4.1.3 Reliabilities of the GCV Section

Reliability analyses were performed to examine the extent to which the items in each section performed as a homogeneous group and the extent to which the items related to other items in the GCV section. The standard errors of measurement were examined to estimate a sort of average of the distribution of error deviations across all the test-takers (Brown, 1996). Grammar and vocabulary sections used Cronbach's alpha reliability estimates for internal consistency while the MC cloze section used the Guttman split half procedure. The reliability estimates and the standard error of measurement are presented in Table 4.5.

Table 4.5
Reliability and Standard Error of Measurement Estimates for the GCV Section

Section	# of Items	Reliability		Standard Error of Measurement	
		Form X	Form Y	Form X	Form Y
Grammar (G)	30	0.69	0.72	2.18	2.12
MC Cloze (C)	20	0.48	0.64	1.80	1.98
Vocabulary (V)	30	0.69	0.70	2.41	2.44
GCV Total	80	0.81	0.85	3.70	3.78

*Reliability estimates for the cloze is a Guttman split-half estimate.

The results showed that all sections yielded alphas of 0.48 or more: grammar (α = 0.69 for Form X, α = 0.72 for Form Y), MC cloze (α = 0.48 for Form X and α = 0.64 for Form Y), and vocabulary (α = 0.69 for Form X, α = 0.70 for Form Y). The values of reliability for grammar and vocabulary sections are similar across forms, but the internal consistency reliability for the MC cloze section is noticeably different across forms. Form X has much lower reliability than Form Y. This indicates that the MC cloze section of Form Y appeared to measure the homogeneous construct within the section more than the items in Form X.

The MC cloze task had the lowest reliability estimate in the GCV section. This implies that, in contrast to the grammar and vocabulary sections, the MC cloze section does not appear to strongly measure a homogeneous construct. It should be noted; however, that the MC cloze and reading sections had smaller number of items (i.e., 20 items each) than the grammar and vocabulary sections (i.e., 30 items each). The small number of items often affects the reliability estimate (Brown, 1996). In order to put the grammar, vocabulary, and MC cloze items on the same scale, the Spearman-Brown Prophecy formula was used (Brown, 1996). When the MC cloze items were scaled to 30 items, the reliability for the cloze section increased to 0.58 for Form X and 0.73 for Form Y.

The standard error of measurement for each section was then examined to estimate an average of the distribution of error deviations. The standard error of measurement for grammar, MC cloze, and vocabulary were 2.18, 1.80, and 2.41, respectively for Form X and 2.12, 1.98, and 2.44 for Form Y. These numbers are similar across forms. However, as mentioned above the MC cloze task only had 20 items on both

forms, and as a result the standard error of measurement was relatively larger than for the 30-item grammar and vocabulary sections for both Form X and Y.

In order to examine the internal consistency reliability for the GCV section of the ECPE, all 80 items were included in the analyses. The results yielded an alpha of 0.81 for Form X and 0.85 for Form Y. The standard error of measurement was 3.70 and 3.78 respectively. The reasonably high alpha of over 0.81 suggests that the items in the exam collectively appear to be measuring the same construct: lexico-grammatical knowledge.

4.1.4 Reliabilities of the GCV Items Based on Theoretical Coding

Reliability analyses for the GCV items based on theoretical coding (i.e., lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning) were performed to examine the extent to which the items in each coding group performed as a homogeneous group, and the extent to which the items related to other items in the coding group. Cronbach's alpha reliability estimates for internal consistency were used for all the coding groups. The reliability estimates and the standard error of measurement are presented in Table 4.6.

Table 4.6
Reliability and Standard Error of Measurement Estimates for the Coded GCV Items

		Reliability		Standard Error of Measurement	
Variables	# of Items	Form X	Form Y	Form X	Form Y
Lexical F.	6	0.21	0.28	0.93	1.11
Morph F.	20	0.61	0.64	1.76	1.72
Cohesive F.	8	0.32	0.38	1.15	1.01
Lexical M.	43	0.71	0.76	2.81	2.91
Cohesive M.	3	0.19	0.25	0.72	0.78

The results showed a variety of reliability estimates. Morphosyntactic form and lexical meaning variables yielded moderately high alphas while lexical form, cohesive form, and cohesive meaning yielded low alphas. It should be noted; however, that the cohesive form, lexical form, and cohesive form have far fewer items than morphosyntactic form and lexical meaning. If the number of items were consistent across the variables, the reliability estimates may have been more consistent.

When the reliability estimates across forms are compared, the reliability estimates for coding variables in Form Y are consistently higher than those of Form X. However, the differences are marginal; hence, no strong conclusions can be drawn from the observation at this point.

4.2 Reading Section

The reading sections of Form X and Form Y were then analyzed based on the responses of 33,662 and 32,473 test-takers respectively. Descriptive statistics and reliability analyses were performed. The reading section consisted of four passages per test form and each passage contained five questions. Therefore, there were a total of twenty questions per test form. All the passages were written as newspaper or magazine articles, and the topics varied from history to natural science.

4.2.1 Section-Level Distribution of the Reading Section

The distribution of the reading section as a whole was examined. Table 4.7 presents the summary distributions of the reading section for Form X and Form Y. The mean for the reading section was 16.58 and 15.54 respectively based on 20 items,

suggesting the Form X was a little easier than Form Y. The standard deviation was 2.98 for Form X and 2.86 for Form Y, which indicates there was not much difference in the dispersion of the data. Skewness and kurtosis were within the acceptable limits for the reading section.

Table 4.7
Distributions of the Reading Section

	# of	Mean		Stdv		Skewness		Kurtosis	
Section	Items	X	Y	X	Y	X	Y	X	Y
Reading	20	16.58	15.54	2.98	2.86	-1.17	-0.88	1.38	0.84

In order to compare means of the variables across Form X and Form Y, t-test was calculated (see Table 4.8). The results show that the differences in means of the reading section were significant. Under the assumption that the test-taking populations for the two forms are identical, this would mean the test forms are at least somewhat different. However, it should again be noted that even a numerically small difference in the means can be significantly different, considering the large sample size of this data. The numerical difference in test means (1.04 points) here is roughly equal to about one-third of a standard deviation in the underlying test scores.

Table 4.8
T-test Results for the Reading Section

	t-test for equality means				
Section	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Reading	45.81	66133	0.00	1.04	0.02

4.2.2 Distribution of the Reading Items Based on Theoretical Coding

Following the section-level distribution analysis, the test items were grouped together into four conceptual categories based on the reading skills: the ability to identify main ideas, the ability to identify details, the ability to understand the meaning of vocabulary in context, and the ability to identify inferences. These categories were based on the theoretical coding described in Chapter 2. Table 4.9 presents the summary distributions of the four variables for Form X and Y.

The mean for the main idea items was 1.48 and 1.44 respectively based on two items. This indicates that the main idea items measured in Form X and Form Y were almost the same in terms of difficulty. The standard deviation was 0.62 for both forms, which indicates there was no difference in the dispersion of the data. Skewness and kurtosis were within the acceptable limits for the main idea items.

The mean for the detail items was 9.10 for Form X and 8.57 for Form Y. This indicates that the detail items were slightly easier on Form X. The standard deviation was 1.79 for Form X and 1.83 for Form Y, which indicates there was not much difference in the dispersion of the data. Skewness and kurtosis were within the acceptable limits for the detail items.

The mean for vocabulary item was 0.87 for Form X and 0.53 for Form Y, which suggests that Form X was much easier than Form Y. The standard deviation was 0.34 for Form X and 0.50 for Form Y, which indicates there was more dispersion for Form Y. Skewness and kurtosis were within the acceptable limits for the vocabulary item.

The mean for the inference items was 5.14 for Form X and 5.00 for Form Y, suggesting that the difficulty level of the inference items were almost the same for Form

X and Form Y. The standard deviation was 1.09 for Form X and 1.07 for Form Y, which indicates there was not much difference in the dispersion of the data. Skewness and kurtosis were within the acceptable limits for the inference items.

Table 4.9
Distributions of the Reading Items Based on Theoretical Coding

Variables	# of items	Mean		Std Dev		Skewness		Kurtosis	
		X	Y	X	Y	X	Y	X	Y
Main Ideas	2	1.48	1.44	0.62	0.62	-0.77	-0.63	-0.42	-0.55
Details	11	9.10	8.57	1.79	1.83	-1.07	-0.83	0.92	0.49
Voc. in context	1	0.87	0.53	0.34	0.50	-2.19	-0.12	2.78	-1.99
Inferences	6	5.14	5.00	1.09	1.07	-1.40	-1.11	1.80	1.11

In order to compare means across Form X and Form Y, t-test was performed. Table 4.10 presents the summary of the t-test. The t value for the vocabulary in context variable is extremely high because the mean difference was noticeably large. The results show that the differences in means for the reading section were significant, which indicates that the variable means on Form X and Form Y were different. Under the assumption that the test-taking populations for the two forms are identical, this would mean the forms are at least somewhat different. It should again be noted; however, that even numerically small differences in the means may be significantly different considering the large sample size of this data. Therefore, it would again be premature to conclude that Form X and Form Y have different underlying structures based on the results of the t-test.

Table 4.10
T-test Results for the Reading Items Based on Theoretical Coding

Variables	t-test for equality means				
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Main Ideas	6.58	66133	0.00	0.03	0.01
Details	37.29	66133	0.00	0.52	0.01
Voc. in context	102.50	66133	0.00	0.34	0.00
Inferences	16.68	66133	0.00	0.14	0.01

4.2.3 Reliabilities of the Reading Section

Following the distribution analyses, reliability analyses were performed to examine the extent to which the items in each section performed as a homogeneous group and the extent to which the items related to other items in the reading section. The standard error of measurement was computed to further examine the score distributions across test-takers (Brown, 1996). The reliability estimates and the standard error of measurement are presented in Table 4.11.

Table 4.11
Reliability and Standard Error of Measurement Estimates for the Reading Section

Section	# of Items	Reliability		Standard Error of Measurement	
		Form X	Form Y	Form X	Form Y
Reading	20	0.74	0.65	1.50	1.69

The results showed that the reliability of Form X ($\alpha = 0.74$) was slightly higher than that of Form Y ($\alpha = 0.65$), indicating the reading section of Form X appeared to measure the same construct within the section more reliably than did the items in Form Y. The standard error of measurement was 1.50 for Form X and 1.69 for Form Y. The difference in the standard error of measurement across forms appears to be marginal.

4.2.4 Reliabilities of the Reading Items Based on Theoretical Coding

Reliability analyses for the reading items based on theoretical coding into four conceptual categories were performed to examine the extent to which the items in each coding variable performed as a homogeneous group, and the extent to which the items related to other items in the coding group. Cronbach's alpha reliability estimates for internal consistency were used for all the coding variables. The reliability estimates and the standard error of measurement are presented in Table 4.12.

Table 4.12
Reliability and Standard Error of Measurement Estimates for the Coded Reading Items

Variables	# of Items	Reliability		Standard Error of Measurement	
		Form X	Form Y	Form X	Form Y
Main Idea	2	0.20	0.06	0.55	0.85
Detail	11	0.56	0.52	1.19	1.27
Voc. in context	1	N/A	N/A	N/A	N/A
Inference	6	0.55	0.39	0.73	0.86

The results showed a variety of reliability estimates. The main idea items generated extremely low alphas ($\alpha = 0.20$ for From X, $\alpha = 0.06$ for Form Y), likely because they were only two items in the group. A small number of items within the group can provide insufficient variability in the items and produce low reliability.

The detail items yielded moderate alphas across forms ($\alpha = 0.56$ for From X, $\alpha = 0.52$ for Form Y). The vocabulary items did not provide reliability estimates as there was only one item in the group. The inference items yielded an alpha of 0.55 for Form X and 0.39 for Form Y. This indicates that inference items in Form X appeared to measure homogeneous construct within the group more reliably than the items in Form Y.

The standard error of measurement for each group was then examined. The standard error of measurement for the main idea items was 0.55 for Form X and 0.85 for

Form Y, indicating that Form Y had a much higher error. By looking at the extremely low reliability for Form Y ($\alpha = 0.06$), it is expected to have higher standard error of measurement. The standard error of measurement for the detail items was 1.19 for Form X and 1.27 for Form Y. These numbers show that the error was not much different for the detail items. The standard error of measurement for the inference items was 0.73 for Form X and 0.86 for Form Y. These numbers show that the error was a little higher for Form Y. Because Form Y ($\alpha = 0.39$) has a lower reliability than Form X ($\alpha = 0.55$), it is expected for Form Y to have a higher standard error of measurement. No reliability estimates were available for the vocabulary in context items, so the standard error of measurement could not be calculated for this group.

4.3 Summary

In this chapter, the results of the section-level analyses performed on the GCVR section of Form X and Form Y were examined. More specifically, the distributions and reliability estimates were discussed as a whole test. In the following chapter, the results of the confirmatory factor analyses for theoretically plausible structure models of the GCVR section for Form X and Form Y are discussed.

Chapter V

SINGLE-GROUP ANALYSES

Based on the preliminary analyses and the substantive literature on lexico-grammatical knowledge and reading ability, underlying trait structures of the GCVR section of ECPE was postulated. This chapter investigates the fit of theoretically plausible models of the GCVR section by performing a series of separate confirmatory factor analyses.

The chapter begins with the examination of the underlying trait structure of the GCV section. Then, the underlying trait structure of the reading section is investigated. The GCV section and the reading section are modeled together in the last part of the chapter.

5.1 Confirmatory Factor Analysis of the GCV Section

In this section, hypothesized models of the GCV section of Form X and Form Y are analyzed using CFA. The section is divided into two parts: the first section examines the model of the GCV section for Form X and the second section for Form Y.

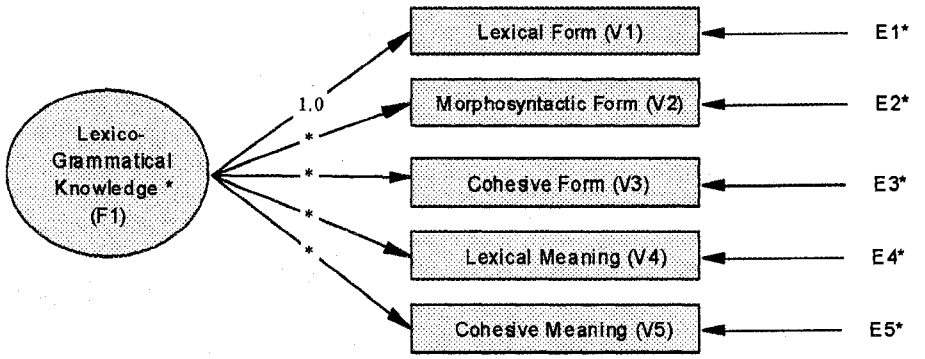
5.1.1 Testing the Factorial Validity of the GCV Section: Form X

The hypothesized model: Model 5.1 (Initial Model of the GCV Section of Form X)

Based on the substantive theory, the GCV section was represented schematically as a one-factor model of lexico-grammatical knowledge. This initially-hypothesized

model is presented in Figure 5.1. It contained one factor (lexico-grammatical knowledge) and five observed variables (lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning, denoted V1-V5 respectively). Each observed variable was hypothesized to load on lexico-grammatical knowledge. The error terms associated with the observed variables (E1 through E5) were postulated to be uncorrelated.

Figure 5.1
Initially-Hypothesized Model for the GCV Section for Form X: Model 5.1



* = freely estimated

Model 5.1 is a first-order confirmatory factor analysis designed to test the underlying constructs of lexico-grammatical knowledge measured by the GCV section of ECPE. Given the exploratory nature of the present study, the current investigation was not limited to a simple confirmation or rejection of this initial model. Rather, the relationships among the variables were explored with the goal of producing the best fitting and most substantively meaningful model. After examining the results for Model 5.1, the study proceeds to identify additional specifications and improvements if necessary.

The results for Model 5.1 (Initial Model of the GCV Section of Form X)

Before investigating how well Model 5.1 fits the data, multivariate normality assumptions were examined. With regard to multivariate kurtosis, the data produced a Mardia's coefficient of 0.97 with a normalized estimate of 10.60. The normalized estimate was beyond Bentler's (2006) suggested range of +3 to -3; hence, it indicated multivariate non-normality. Based on the results, maximum likelihood robust estimation method was used to account for non-normality in the data (Bentler, 2006). When the robust estimation is used, a robust chi-square statistic called the Satorra-Bentler scaled statistic (Satorra & Bentler, 1994) and robust standard errors (Bentler & Dijkstra, 1985) are provided in the EQS output. Both of these values are corrected for non-normality in large samples. The current study used Satorra-Bentler scaled chi-square statistic (S-Bχ^2) when there was a multivariately non-normal data.

Following the examination of Mardia's coefficient, possible multivariate outliers in the data were checked. EQS listed five cases that contributed most to the normalized multivariate kurtosis. The size of these estimates was compared to one another to see if any of them were strikingly different from the others. The estimates were all similar; suggesting these five cases were not in fact multivariate outliers.

Checking Mahalanobis distance statistic is another way of examining multivariate outliers (DeCarlo, 1997). It measures the multivariate distance between scores of an individual case and the sample means. The Mahalanobis distance would be zero, if the score of a particular case equals its respective mean. If the Mahalanobis distance statistic for a specific case is significantly different from zero, that case may be considered as a multivariate outlier. To investigate the presence of possible multivariate outliers, the

Mahalanobis distance statistics were calculated for all cases using an SPSS macro called *normtest* (DeCarlo, 1997). The critical value for a single multivariate outlier is at the 0.05 level. There were some cases with the critical value over 0.05; however, it did not make much difference in the value of normalized estimate when those cases were deleted. This is due to the large number of cases in the data (N>32,000). Therefore, none of the possible outliers were deleted in the current study.

After the multivariate outliers were checked, all other statistical assumptions of the estimation procedure, such as identification and number of iterations for conversion were examined, and there were no violations in the data.

The next step was to assess the hypothesized model to determine to what extent the model fit the sample data. The results produced a Satorra-Bentler chi-square (S-Bχ^2) of 223.95 with 5 degrees of freedom (df) ($p<0.01$). Although the chi-square/df ratio was beyond the recommended value of 3, the chi-square likelihood ratio test is known for being sensitive to sample size (Bentler & Bonett, 1980). Hence, other goodness of fit indices were examined. Model 5.1 produced a CFI (Comparative Fit Index) of 0.990 and a NNFI (Bentler-Bonnet Non-Normed Fit Index) of 0.981. A value for the RMSEA (Root Mean Square Error of Approximation) was 0.036, which was within the acceptable limit. In summary, all these fit indices suggested that Model 5.1 may be an adequate representation of the data. These results are presented in Table 5.1.

Table 5.1
Results for the Initially-Hypothesized Model for the GCV Section for Form X: Model 5.1

Multivariate Kurtosis	
Mardia's coefficient (G2, P)	0.97
Normalized estimate	10.60
Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	223.95
Degrees of freedom	5
Probability value for the chi-square statistic	0.0000
Fit Indices	
Comparative Fit Index (CFI)	0.990
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.981
Root Mean-Square Error of Approximation (RMSEA)	0.036

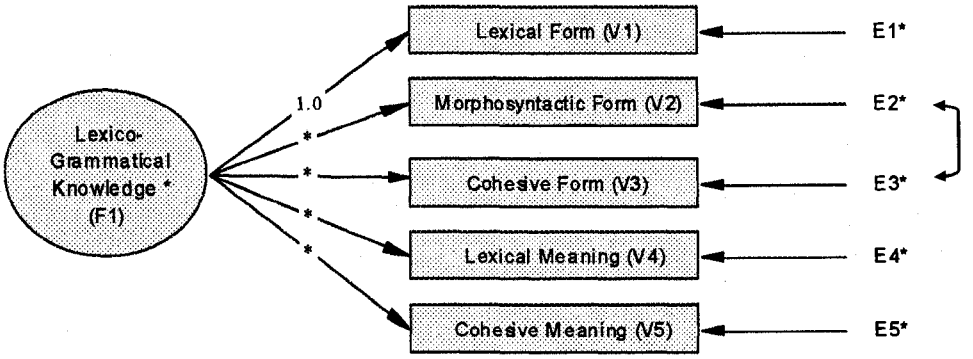
Although Model 5.1 represented the data well, a series of post hoc fitting procedures (e.g., the LM and Wald tests) were performed to see whether there was a better fitting model for the sample data.

The hypothesized model: Model 5.2 (Revised Model of the GCV Section of Form X)

The LM test suggested that the error terms for morphosyntactic form and cohesive form should be correlated. It is substantively reasonable to correlate these two error terms as both variables measure the form dimension of lexico-grammatical knowledge. Hence, Model 5.2 was built based on both a substantive and statistical point of view.

Model 5.2 is summarized as follows. Lexico-grammatical knowledge is measured by five observed variables (lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning). Then, the error terms associated with morphosyntactic form and cohesive form were hypothesized to be correlated. A diagrammatic representation of an alternative model is shown in Figure 5.2.

Figure 5.2
Revised Model for the GCV Section for Form X: Model 5.2



* = freely estimated

The results for Model 5.2 (Revised Model of the GCV Section of Form X)

Based on the sample statistics, all five variables showed satisfactory skewness and kurtosis values. With regard to multivariate kurtosis, the data again produced a Mardia's coefficient of 0.97 with a normalized estimate of 10.60. The normalized estimate was beyond Bentler's (2006) suggested range of +3 to -3; hence, it indicated multivariate non-normality. Based on the results, maximum likelihood robust estimation method was used to account for non-normality in the data (Bentler, 2006).

With respect to goodness of fit, Model 5.2 produced a Satorra-Bentler chi-square (S-Bχ^2) of 203.62 with 4 degrees of freedom, representing a drop in overall chi-square (ΔS-Bχ^2 (1) = 20.33) from the initially-hypothesized model. This decrease in S-Bχ^2 exhibited an improvement in goodness of fit. Consistent with this statistical improvement, the CFI (0.992) also reflected an improvement in model data fit (Δ = 0.002). The RMSEA dropped from 0.036 in Model 5.1 to 0.034 in Model 5.2 (Δ = -0.002). In

summary, all these fit indices and the improvement in the S-Bχ^2 statistic suggested that Model 5.2 is a better representation of the data. These results are presented in Table 5.2.

Table 5.2
Results for the Revised Model for the GCV Section for Form X: Model 5.2

Multivariate Kurtosis	
Mardia's coefficient (G2, P)	0.97
Normalized estimate	10.60
Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	203.62
Degrees of freedom	4
Probability value for the chi-square statistic	0.0000
Fit Indices	
Comparative Fit Index (CFI)	0.992
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.981
Root Mean-Square Error of Approximation (RMSEA)	0.034

Given that Model 5.2 represented the data well, the feasibility of the individual parameter estimates was checked. All estimates were found to be reasonable and statistically significant at the 0.05 level. The variances and covariances of all the independent variables were also statistically significant at the 0.05 level. This suggests that the underlying factors were well measured by the observed variables and that these variables were measuring extant lexico-grammatical knowledge.

As shown in Table 5.3, the standardized solution of factor loadings for Model 5.2 were all within the acceptable limits and ranged from a moderately low 0.38 for cohesive meaning to a high 0.71 for morphosyntactic form. Many of the items in the grammar section were measuring morphosyntactic form; hence, it is reasonable to see a high parameter estimate on the morphosyntactic form variable. The second highest parameter

estimate value was lexical meaning (0.64). This is also expected as many of the items in the vocabulary section were measuring lexical meaning.

In sum, the GCV section of Form X was explained by lexico-grammatical knowledge. Lexico-grammatical knowledge was well-measured by the test-takers' ability to identify morphosyntactic form and lexical meaning in various test items. Both variables displayed a relatively strong, significant (at 0.05 level) association with lexico-grammatical knowledge.

Neither the standardized estimates nor the errors were found to be outside the acceptable range. The proportion of the variances of the measured variables accounted for by their hypothesized latent variable (i.e., R-squared) ranged from 0.142 to 0.510. More specifically, no estimates of variances were near zero or negative values.

Table 5.3
Parameter Estimates for the Revised Model for the GCV Section for Form X: Model 5.2

STANDARDIZED SOLUTION:									R-SQUARED
Lex F	=	V1	=	0.49	F1	+	0.87	E1	0.242
Morph F	=	V2	=	0.71	*F1	+	0.70	E2	0.510
Coh F	=	V3	=	0.59	*F1	+	0.81	E3	0.351
Lex M	=	V4	=	0.64	*F1	+	0.77	E4	0.415
Coh M	=	V5	=	0.38	*F1	+	0.93	E5	0.142

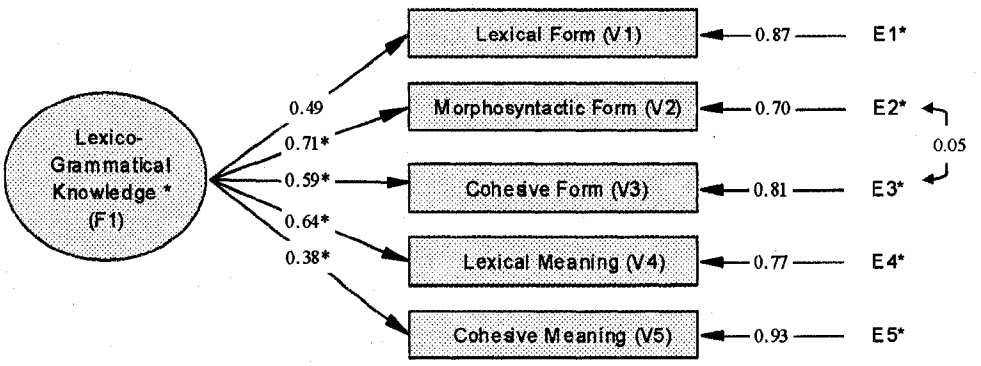
F1 = Lexico-grammatical Knowledge
Lex F = Lexical Form
Morph F = Morphosyntactic Form
Coh F = Cohesive Form
Lex M = Lexical Meaning
Coh M = Cohesive Meaning
* = freely estimated

Figure 5.3 presents a diagrammatic representation of Model 5.2 in which the standardized parameter estimates are indicated. An inspection of Model 5.2 shows that the GCV section for Form X is represented by a first-order factor called lexico-

grammatical knowledge measured by five observed variables. This model supports the work of Larsen-Freeman (1991) and Purpura (1999, 2004) in their claims that grammatical knowledge consists of form and meaning. In the current study, form and meaning were represented by lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning.

With regard to error terms, Model 5.2 produced one pair of correlated error terms. The correlation was 0.05, indicating that there was some redundant content being measured between morphosyntactic form and cohesive form. The redundancy is appropriate as both morphosyntactic form and cohesive form measure the form dimension of lexico-grammatical knowledge.

Figure 5.3
Revised Model for the GCV Section for Form X with
Standardized Parameter Estimates: Model 5.2



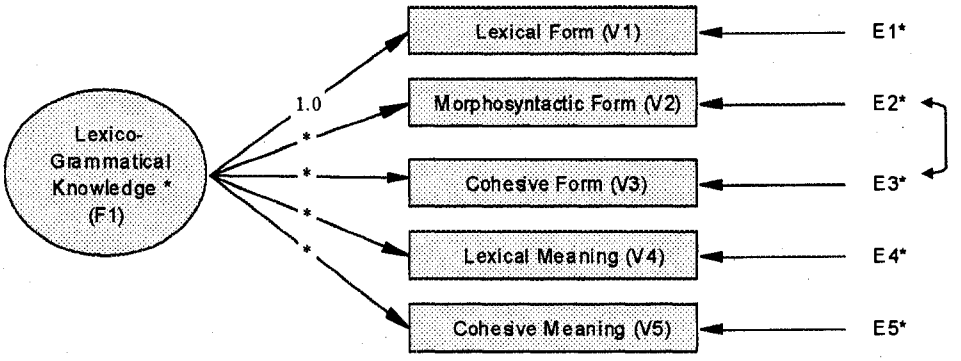
5.1.2 Testing the Factorial Validity of the GCV Section: Form Y

The hypothesized model: Model 5.3

Having selected the final model for Form X (Model 5.2), the fit for the same model was tested for Form Y. Model 5.3 is a hypothesized model for Form Y, which is

presented in Figure 5.4. In this model, lexico-grammatical knowledge is measured by five observed variables (lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning). Then, the error terms associated with morphosyntactic form and cohesive form were hypothesized to be correlated. A diagrammatic representation of an alternative model is shown in Figure 5.4.

Figure 5.4
Hypothesized Model for the GCV Section for Form Y: Model 5.3



The results for Model 5.3 (Hypothesized Model of the GCV Section of Form Y)

Based on the sample statistics, all five variables showed satisfactory skewness and kurtosis values. With regard to multivariate kurtosis, the data again produced a Mardia's coefficient of -0.71 with a normalized estimate of -7.63. The normalized estimate was beyond Bentler's (2006) suggested range of +3 to -3; hence, it indicated multivariate non-normality. Based on the results, maximum likelihood robust estimation method was used to account for non-normality in the data (Bentler, 2006).

With respect to goodness of fit, Model 5.3 produced a Satorra-Bentler chi-square (S-Bχ^2) of 169.97 with 4 degrees of freedom. It produced a CFI of 0.998 and a NFI of

0.988. A value of the RMSEA was within the acceptable limit. Overall, the goodness of fit indices suggest that Model 5.3 is a good representation of the data. These results are presented in Table 5.4.

Table 5.4
Results for the Hypothesized Model for the GCV Section for Form Y: Model 5.3

Multivariate Kurtosis	
Mardia's coefficient (G2, P)	-0.71
Normalized estimate	-7.63
Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	169.97
Degrees of freedom	4
Probability value for the chi-square statistic	0.0000
Fit Indices	
Comparative Fit Index (CFI)	0.998
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.988
Root Mean-Square Error of Approximation (RMSEA)	0.036

Given that Model 5.3 represented the data well, the feasibility of the individual parameter estimates was checked. All estimates were found to be reasonable and statistically significant at the 0.05 level. The variances and covariances of all the independent variables were also statistically significant at the 0.05 level. This suggests that the underlying factors were well measured by the observed variables and that these variables were measuring extant lexico-grammatical knowledge.

As shown in Table 5.5 the standardized solution of factor loadings for Model 5.3 were all within the acceptable limits and ranged from a moderately low 0.37 for cohesive meaning to a high 0.81 for lexical meaning. Many of the items in the GCV section were measuring lexical meaning; hence, it is reasonable to see a high parameter estimate on the

lexical meaning variable. The second highest parameter estimate value was morphosyntactic form (0.70). This is also expected as many of the items in the GCV section were measuring morphosyntactic form.

In sum, the GCV section of Form Y was explained by lexico-grammatical knowledge. Lexico-grammatical knowledge was well-measured by the test-takers' ability to identify morphosyntactic form and lexical meaning in various test items. Both variables displayed a relatively strong, significant (at 0.05 level) association with lexico-grammatical knowledge.

Neither the standardized estimates nor the errors were found to be outside the acceptable range. The proportion of the variances of the measured variables accounted for by their hypothesized latent variables (i.e., R-squared) ranged from 0.135 to 0.656. More specifically, no estimates of variances were near zero or negative values.

Table 5.5
Parameter Estimates for the Hypothesized Model for the GCV Section for Form Y: Model 5.3

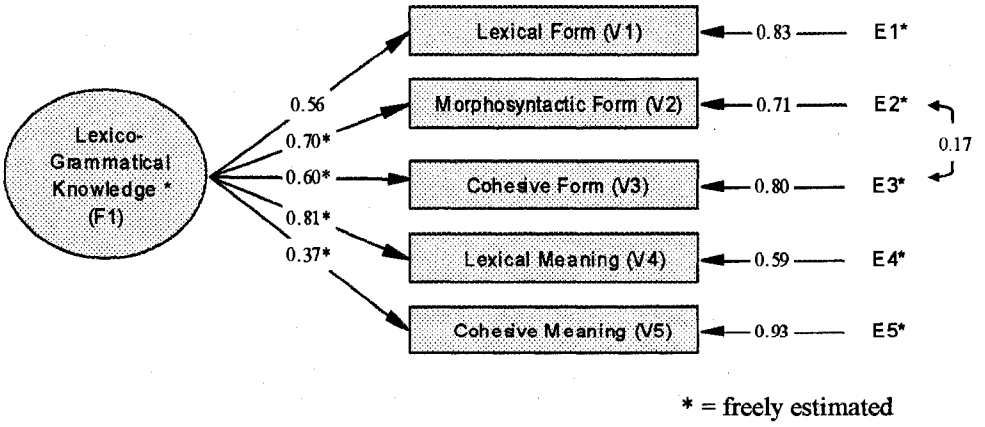
STANDARDIZED SOLUTION:									R-SQUARED
Lex F	=	V1	=	0.56	F1	+	0.83	E1	0.312
Morph F	=	V2	=	0.70	*F1	+	0.71	E2	0.490
Coh F	=	V3	=	0.60	*F1	+	0.80	E3	0.359
Lex M	=	V4	=	0.81	*F1	+	0.59	E4	0.656
Coh M	=	V5	=	0.37	*F1	+	0.93	E5	0.135

F1 = Lexico-grammatical Knowledge
Lex F = Lexical Form
Morph F = Morphosyntactic Form
Coh F = Cohesive Form
Lex M = Lexical Meaning
Coh M = Cohesive Meaning
* = freely estimated

Figure 5.5 presents a diagrammatic representation of Model 5.3 in which the standardized parameter estimates are indicated. An inspection of Model 5.3 shows that the GCV section for Form Y is represented by a first-order factor called lexico-grammatical knowledge measured by five observed variables. This model again supports the work of Larsen-Freeman (1991) and Purpura (1999, 2004) in their claims that grammatical knowledge consists of form and meaning. In the current study, form and meaning were represented by lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning.

With regard to error terms, Model 5.3 produced one pair of correlated error terms. The correlation was 0.17, indicating that there was some redundant content being measured between morphosyntactic form and cohesive form. The redundancy is appropriate as both morphosyntactic form and cohesive form measure the form dimension of lexico-grammatical knowledge.

Figure 5.5
Hypothesized Model for the GCV Section for Form Y with Standardized Parameter Estimates: Model 5.3



In summary, the model of lexico-grammatical knowledge fit the data well for both Form X and Form Y. The factor loadings were very similar between the two forms with the cohesive meaning loading being the lowest, and the morphosyntactic form and lexical meaning being the two highest loadings for both Form X and Form Y. The results suggest that the underlying constructs of Form X and Form Y for the GCV section appear to be comparable to one another.

5.2 Confirmatory Factor Analysis of the Reading Section

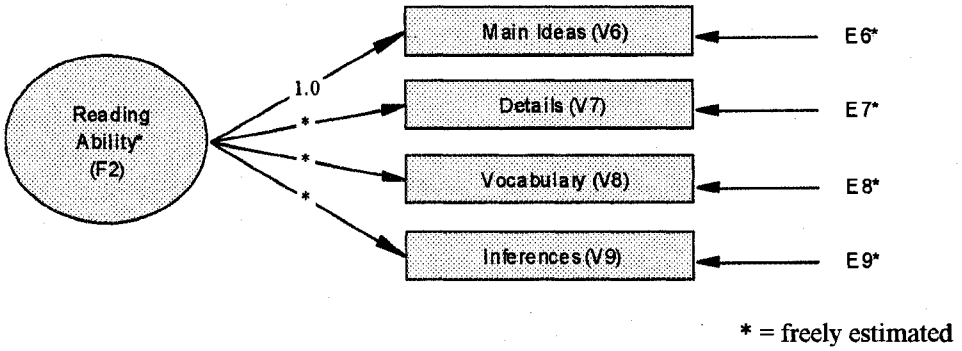
In this section, hypothesized models of the reading section of Form X and Form Y are analyzed using confirmatory factor analysis (CFA). The section is divided into two parts: the first section examines the model of the reading section for Form X and the second section for Form Y.

5.2.1 Testing the Factorial Validity of the Reading Section: Form X

The hypothesized model: Model 5.4

Based on the substantive theory on reading skills, the reading section was represented schematically as a one-factor model of reading ability. This initially-hypothesized model is presented in Figure 5.6. It contained one factor (reading ability) and four observed variables (main ideas, details, vocabulary in context, and inferences, denoted V6-V9 respectively). Each observed variable was hypothesized to load on reading ability. The error terms associated with the observed variables (E6 through E9) were postulated to be uncorrelated.

Figure 5.6
Initially-Hypothesized Model for the Reading Section for Form X: Model 5.4



Model 5.4 is a first-order confirmatory factor analysis designed to test the underlying constructs of reading ability measured by the reading section of ECPE. Given the exploratory nature of the present study, the current investigation was not limited to a simple confirmation or rejection of this initial model. Rather, the relationships among the variables were explored with the goal of producing the best fitting and most substantively meaningful model. After examining the results for Model 5.4, the study proceeds to identify additional specifications and improvements if necessary.

The results for Model 5.4 (Initial Model of the Reading Section of Form X)

Before exploring how well the hypothesized model fits the data, multivariate normality assumptions were examined. The data produced a Mardia's coefficient of 8.29 with a normalized estimate of 109.77, suggesting multivariate non-normality. Subsequently, possible multivariate outliers in the data were checked. Among the five cases listed in the EQS program, the case number 30882 appeared strikingly different from the others. Therefore, it was deleted from the analysis. When the case number

30882 was deleted, a Mardia's coefficient went down from 8.29 to 8.27 (Δ-0.02) and a normalized estimate went down from 109.77 to 109.55 (Δ-0.22).

Because the changes in the values were so marginal that the deletion of the case did not seem to affect the outcome of the results, the current study used the data without the case deleted. Due to the multivariately non-normality in the data, maximum likelihood robust estimation method was used to account for non-normality in the data (Bentler, 2006).

After examining the multivariate kurtosis and outliers, other statistical assumptions of the estimation procedure, such as identification and number of iterations for conversion, were examined. There were no violations found; hence, the fit of the hypothesized model was then tested.

The goodness of fit index for the initially-hypothesized one-factor model of reading ability produced a Satorra-Bentler chi-square (S-Bχ^2) of 66.30 with 2 degrees of freedom (df). Although the chi-square/df ratio was beyond the recommended value of 3, the chi-square likelihood ratio test is known for being sensitive to sample size (Bentler & Bonett, 1980). Hence, other goodness of fit indices were examined. Model 5.4 produced a CFI of 0.995 and a NNFI of 0.984. A value for the RMSEA was 0.031, which was within the acceptable range. These results indicate a good fit for the data, provided the individual parameter estimates are viable and statistically significant. Table 5.6 presents the goodness of fit summary for the initially-hypothesized model for the reading section of Form X.

Table 5.6
Results for the Initially-Hypothesized Model for
the Reading Section for Form X: Model 5.4

Multivariate Kurtosis	
Mardia's coefficient (G2, P)	8.29
Normalized estimate	109.77
Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	66.30
Degrees of freedom	2
Probability value for the chi-square statistic	0.0000
Fit Indices	
Comparative Fit Index (CFI)	0.995
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.984
Root Mean-Square Error of Approximation (RMSEA)	0.031

Examining the feasibility of the individual parameters, the estimates of the measurement equation with standard errors and test statistics for the unstandardized solution were reasonable and statistically significant at the 0.05 level. The variances of the error terms were also statistically significant. This suggests that the underlying factors were well measured by the observed variables and that these variables were measuring extant reading ability.

Turning to the standardized solution presented in Table 5.7, the factor loadings for Model 5.4 were all within the acceptable limits. The loadings ranged from a moderately low 0.33 for vocabulary in context to a high 0.79 for details. It is reasonable to see a low loading on the vocabulary in context variable as there was only one item measuring the ability to recognize the appropriate meaning of words in the given context. On the other hand, there were many items (eleven out of twenty) measuring the ability to recognize specific information explicitly stated in the text (i.e., detail items). Therefore, it was

expected to see a high loading on the details variable. In other words, reading ability was well-measured by the test-takers' ability to recognize specific information explicitly stated in the text. Also, it was well-measured by the test-takers' ability to identify inferences as the inferences variable displayed a relatively strong (0.73) association with reading ability.

Neither the standardized estimates nor the errors were found to be outside the acceptable range. The proportion of the variances of the measured variables accounted for by their hypothesized latent variable (i.e., R-squared) ranged from 0.107 to 0.618. More specifically, no estimates of variances were near zero or negative values.

Table 5.7
Parameter Estimates for the Initially-Hypothesized Model for the Reading Section for Form X: Model 5.4

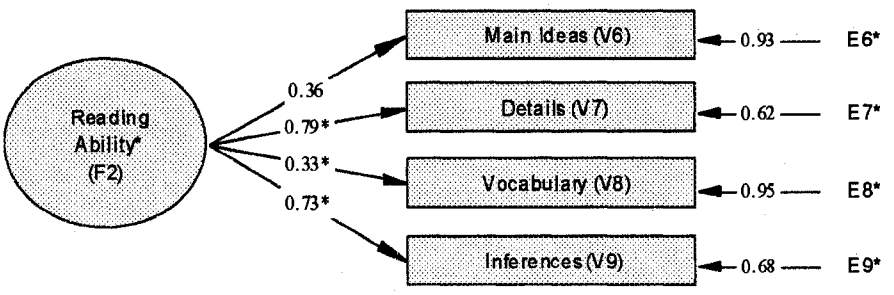
STANDARDIZED SOLUTION:									R-SQUARED
Main	=	V6	=	0.36	F2	+	0.93	E6	0.130
Detail	=	V7	=	0.79	*F2	+	0.62	E7	0.618
Voc	=	V8	=	0.33	*F2	+	0.95	E8	0.107
Infer	=	V9	=	0.73	*F2	+	0.68	E9	0.532

F2 = Reading Ability
Main = Main Ideas
Detail = Details
Voc = Vocabulary in Context
Infer = Inferences
* = freely estimated

Figure 5.7 presents a diagrammatic representation of Model 5.4 in which the standardized parameter estimates are indicated. An inspection of Model 5.4 shows that the reading section for Form X is represented by a first-order factor called reading ability measured by four observed variables. In the current study, reading ability was represented

by the ability to identify main ideas, the ability to identify details, the meaning of vocabulary in context, and the ability to identify inferences.

Figure 5.7
Initially-Hypothesized Model for the Reading Section for Form X with Standardized Parameter Estimates: Model 5.4



* = freely estimated

In summary, Model 5.4 provided evidence for acceptance of the one-factor model of reading ability with a reasonable explanation of the correlations among the observed variables. This finding does not exclude the possibility that reading ability may include other skills. However, the data in the present study support the unidimensional notion of reading ability, where reading ability is represented by the ability to identify main ideas, the ability to identify details, the meaning of vocabulary in context, and the ability to identify inferences.

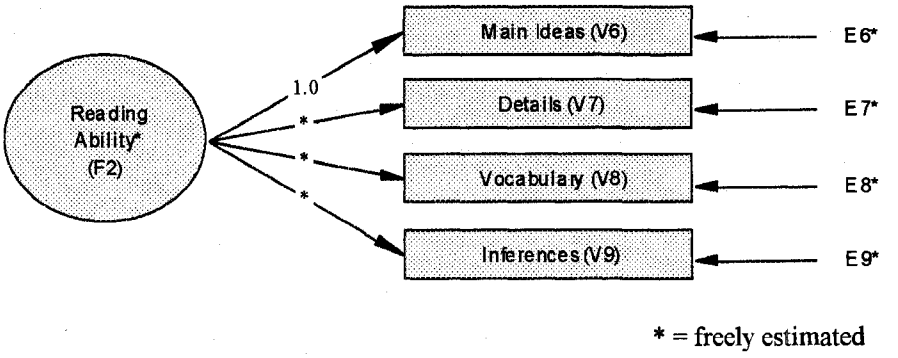
5.2.2 Testing the Factorial Validity of the Reading Section: Form Y

The hypothesized model: Model 5.5

Having selected the final model for Form X (Model 5.4), the fit for the same model was tested for Form Y. Model 5.5 is a hypothesized model for Form Y, which is presented in Figure 5.8. This model contained one factor (reading ability) and four

observed variables (main ideas, details, vocabulary in context, and inferences, denoted V6-V9 respectively). Each observed variable was hypothesized to load on reading ability. The error terms associated with the observed variables (E6 through E9) were postulated to be uncorrelated.

Figure 5.8
Initially-Hypothesized Model for the Reading Section for Form Y: Model 5.5



Model 5.5 is a first-order confirmatory factor analysis designed to test the underlying construct of reading ability measured by the reading section of ECPE. Given the exploratory nature of the present study, the current investigation was not limited to a simple confirmation or rejection of this initial model. Rather, the relationships among the variables were explored with the goal of generating the best fitting and most substantively meaningful model. After examining the results for Model 5.5, the study proceeds to identify additional specifications and improvements if necessary.

The results for Model 5.5 (Initial Model of the Read Section of Form Y)

Prior to exploring the underlying trait structures of Model 5.5, multivariate normality assumptions were examined. The data produced a Mardia's coefficient of -0.14

with a normalized estimate of -1.87, suggesting multivariate normality. Since the data was multivariately normal, a maximum likelihood estimation method was used instead of maximum likelihood robust estimation. All other statistical assumptions of the estimation procedures were checked, and no violations were found.

The goodness of fit index for the initially-hypothesized one-factor model of reading ability produced a chi-square value of 18.73 with 2 degrees of freedom (df), and a CFI of 0.998 and a NNFI of 0.995. A value for the RMSEA was 0.016, which was within the acceptable range. Such high fit indices with low chi-square value suggested that Model 5.5 might be an excellent representation of the data. Table 5.8 presents the goodness of fit summary for the initially-hypothesized model for the reading section of Form Y.

Table 5.8
Results for the Hypothesized Model for
the Reading Section for Form Y: Model 5.5

Multivariate Kurtosis	
Mardia's coefficient (G2, P)	-0.14
Normalized estimate	-1.87
Goodness of Fit Summary Method = Maximum Likelihood	
Chi-square	18.73
Degrees of freedom	2
Probability value for the chi-square statistic	0.0000
Fit Indices	
Comparative Fit Index (CFI)	0.998
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.995
Root Mean-Square Error of Approximation (RMSEA)	0.016

Following the examination of fit indices, the feasibility of the individual parameter estimates was checked, and they were found to be reasonable and statistically

significant at the 0.05 level. The variances of the error terms were also found to be statistically significant. This suggests that the underlying factors were well measured by the observed variables and that these variables were measuring extant reading ability.

As shown in Table 5.9, the standardized solution of factor loadings for Model 5.5 was all within the acceptable limits. The lowest loading was 0.28 for the vocabulary in context variable and the main ideas variable. It is reasonable to see a low loading on these variables as there were two or fewer items measuring these variables. In other words, these variables did not contribute to the construct of reading ability as much as the other variables with higher loadings. The highest loading was 0.69 for the details variable and the second highest was 0.67 for the inferences variable. This suggests that reading ability was well-measured by the test-takers' ability to recognize specific information explicitly stated in the text (i.e., detail items) and the ability to identify inferences.

Neither the standardized estimates nor the errors were found to be outside the acceptable range. The proportion of the variances of the measured variables accounted for by their hypothesized latent variable (i.e., R-squared) ranged from 0.079 to 0.471. More specifically, no estimates of variances were near zero or negative values.

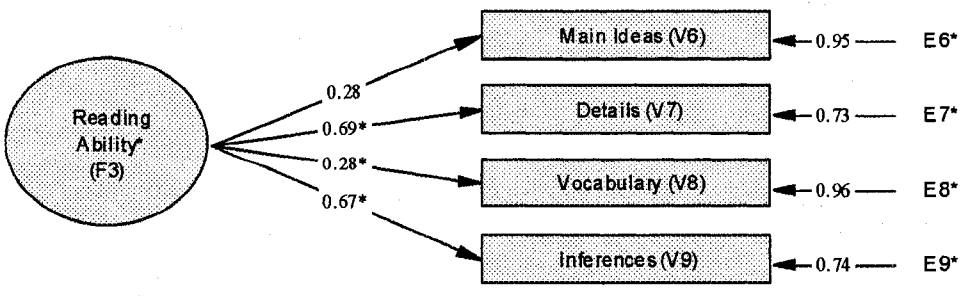
Table 5.9
Parameter Estimates for the Initially-Hypothesized Model for the Reading Section for Form Y: Model 5.5

STANDARDIZED SOLUTION:									R-SQUARED
Main	=	V6	=	0.28	F2	+	0.95	E6	0.079
Detail	=	V7	=	0.69	*F2	+	0.73	E7	0.471
Voc	=	V8	=	0.28	*F2	+	0.96	E8	0.081
Infer	=	V9	=	0.67	*F2	+	0.74	E9	0.446

F2 = Reading Ability
Main = Main Ideas
Detail = Details
Voc = Vocabulary in Context
Infer = Inferences
* = freely estimated

Figure 5.9 presents a diagrammatic representation of Model 5.5 in which the standardized parameter estimates are indicated. An inspection of Model 5.5 shows that the reading section for Form Y is represented by a first-order factor called reading ability measured by four observed variables. In the current study, reading ability was represented by the ability to identify main ideas, the ability to identify details, the meaning of vocabulary in context, and the ability to identify inferences. This model is identical to that of Form X.

Figure 5.9
Initially-Hypothesized Model for the Reading Section for Form Y with Standardized Parameter Estimates: Model 5.5



In summary, Model 5.5 provided evidence for acceptance of the one-factor model of reading ability with a reasonable explanation of the correlations among the observed variables. This finding does not exclude the possibility that reading ability may include other skills. However, the data in the present study support the unidimensional notion of reading ability, where reading ability is represented by the ability to identify main ideas, the ability to identify details, the meaning of vocabulary in context, and the ability to identify inferences.

When the parameter estimates for Form X and Form Y were compared, they showed a similar pattern of loadings. The loadings for the details variable were the highest and the loadings for the vocabulary in context variable were the lowest for both forms. Moreover, the parameter estimates were similar across both forms. These findings suggest that the models for Form X and Form Y may be comparable to one another.

5.3 Confirmatory Factor Analysis of the GCVR Section

The models for the GCV and reading sections were separately tested in the previous parts of this chapter. The next step is to combine these models to examine the underlying trait structures of the GCVR section. More specifically, the model combines the constructs of the lexico-grammatical knowledge and reading ability measured by the GCVR section of ECPE. The following sections are divided into two parts: the first section examines the model of the GCVR section for Form X and the second section for Form Y.

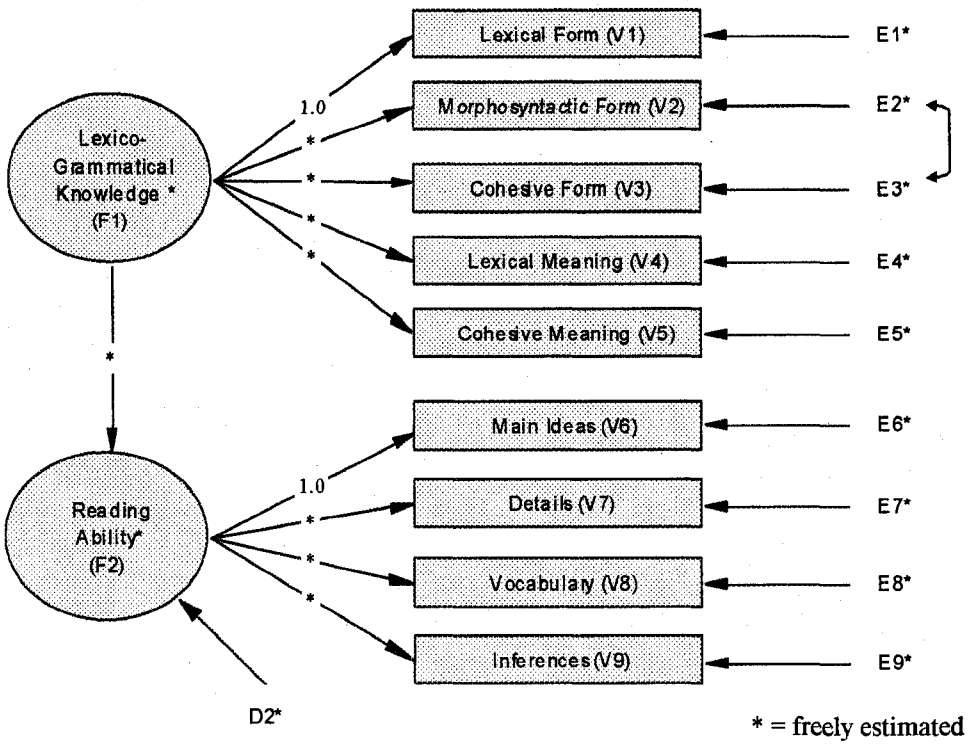
5.3.1 Testing the Factorial Validity of the GCVR Section: Form X

The hypothesized model: Model 5.6

Based on the substantive theory and the results of the separate CFAs for lexico-grammatical knowledge and reading ability, the GCVR section for Form X was represented schematically as a two-factor model of lexico-grammatical knowledge and reading ability. It contained two factors (lexico-grammatical knowledge and reading ability) with nine observed variables (lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning, denoted V1-V5 respectively, and main ideas, details, vocabulary in context, and inferences, denoted V6-9 respectively). Each observed variable was hypothesized to load on only one factor. The error terms associated with the observed variables were postulated to be uncorrelated except the one pair of correlated error terms between morphosyntactic form and cohesive form. This pair of error terms was correlated based on the results from the CFA of the GCV section for Form X (see Model 5.2).

Based on theory and empirical research conducted by Alderson (1993) and Purpura (1999), it was postulated that lexico-grammatical knowledge is a critical linguistic resource for reading ability, and that lexico-grammatical knowledge has a strong effect on reading ability. This was also supported by a series of studies (e.g., Bernhardt & Kamil, 1995; Bossers, 1991; Brisbois, 1995; Carrell, 1991), which concluded that L2 reading is primarily dependent on grammatical ability in the second language. Therefore, the model in the current study includes a unidirectional arrow from lexico-grammatical knowledge to reading ability. This initially-hypothesized model is presented in Figure 5.10.

Figure 5.10
Initially-Hypothesized Model for the GCVR Section for Form X: Model 5.6



Model 5.6 is a first-order confirmatory factor analysis designed to test the multidimensionality of the GCVR section of Form X. More specifically, it tests the hypothesis that the underlying construct of the GCVR section is a multidimensional construct composed of lexico-grammatical knowledge and reading ability. Given the exploratory nature of the present study, the current investigation was not limited to a simple confirmation or rejection of this initial model. Rather, the relationships among the variables were explored with the goal of producing the best fitting and most substantively meaningful model. After examining the results for Model 5.6, the study proceeds to identify additional specifications and improvements if necessary.

Model 5.6 addresses the following research question:

Q1: What is the underlying trait structure of lexico-grammatical knowledge and reading ability as measured by the GCVR section in Form X of the ECPE?

The results for Model 5.6 (Initial Model of the GCVR Section of Form X)

Before investigating how well Model 5.6 fits the data, multivariate normality assumptions were examined. With regard to multivariate kurtosis, the data produced a Mardia's coefficient of 12.95 with a normalized estimate of 84.40. The normalized estimate was beyond Bentler's (2006) suggested range of +3 to -3; hence, it indicated multivariate non-normality.

Subsequently, possible multivariate outliers in the data were checked. Among the five cases listed in the EQS program, the case number 18707 appeared strikingly different from the others. Therefore, it was once deleted from the analysis. When the case number 18707 was deleted, a Mardia's coefficient went down from 12.95 to 12.89 (Δ-0.06) and a normalized estimate went down from 84.40 to 84.05 (Δ-0.35). Because the changes in the values were so marginal that the deletion of the case did not seem to affect the outcome of the results, the current study used the data with no case deleted. The data was multivariately non-normal; hence, maximum likelihood robust estimation method was used to account for non-normality in the data (Bentler, 2006).

The goodness of fit index for the initially-hypothesized two-factor model of lexico-grammatical knowledge and reading ability produced a Satorra-Bentler chi-square (S-Bχ^2) of 681.29 with 25 degrees of freedom (df) ($p<0.01$). Although the chi-square/df ratio was beyond the recommended value of 3, the chi-square likelihood ratio test is

known for being sensitive to sample size (Byrne, 2006). Hence, other goodness of fit indices were examined. Model 5.6 produced a CFI of 0.985 and a NNFI of 0.979. A value for the RMSEA was 0.028, which was within the acceptable range. These results indicate a good fit for the data, provided the individual parameter estimates are viable and statistically significant. Table 5.10 presents the goodness of fit summary for the initially-hypothesized model for the GCVR section of Form X.

Table 5.10
Results for the Initially-Hypothesized Model for the GCVR Section for Form X: Model 5.6

Multivariate Kurtosis	
Mardia's coefficient (G2, P)	12.95
Normalized estimate	84.40
Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	681.29
Degrees of freedom	25
Probability value for the chi-square statistic	0.0000
Fit Indices	
Comparative Fit Index (CFI)	0.985
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.979
Root Mean-Square Error of Approximation (RMSEA)	0.028

These fit statistics provided strong evidence for acceptance of Model 5.6. A series of post hoc fitting procedures (e.g., the LM and Wald tests) were performed to see whether there was a better fitting model for the sample data. However, none of the alternative models indicated by these procedures were theoretically plausible; hence, the initially-hypothesized model (Model 5.6) was treated as the best fitting model.

In examining the measurement equations of the individual parameter estimates, all estimates were found to be reasonable and statistically significant at the 0.05 level.

The variances and covariances of all the independent variables were also statistically significant at the 0.05 level. As shown in Table 5.11, the standardized solution of factor loadings for Model 5.6 were all within the acceptable limits and ranged from a low moderate 0.33 for vocabulary in context to a high 0.80 for details. The fairly high loading of 0.68 from lexico-grammatical knowledge to reading ability provided evidence that lexico-grammatical knowledge had an effect on reading ability in this test.

The factor loading patterns were similar to the original model for the GCV section. In Model 5.2, where the GCV section was modeled by itself, the two highest loading variables measuring lexico-grammatical knowledge were morphosyntactic form and lexical meaning. In the current model (Model 5.6), the highest loading variables were the same.

A pattern of loadings for the reading section also did not change from the original reading ability model (Model 5.4). The variable measuring the ability to identify details was again contributing the most to the construct of reading ability (0.80) and the variable measuring the meaning of vocabulary in context was again contributing the least (0.33).

With regard to standardized estimates, none of the estimates or the errors was found to be outside the acceptable range. The proportion of the variances of the measured variables accounted for by their hypothesized latent variable (i.e., R-squared) ranged from 0.107 to 0.642. More specifically, no estimates of variances were near zero or negative values.

Table 5.11
Parameter Estimates for the Revised Model for
the GCVR Section for Form X: Model 5.6

STANDARDIZED SOLUTION:								R-SQUARED
Lex F	= V1	=	0.48	F1	+	0.88	E1	0.231
Morph F	= V2	=	0.66	*F1	+	0.74	E2	0.456
Coh F	= V3	=	0.56	*F1	+	0.83	E3	0.316
Lex M	= V4	=	0.68	*F1	+	0.73	E4	0.466
Coh M	= V5	=	0.40	*F1	+	0.92	E5	0.157
Main Ideas	= V6	=	0.37	F2	+	0.93	E6	0.137
Details	= V7	=	0.80	*F2	+	0.60	E7	0.642
Voc in Context	= V8	=	0.33	*F2	+	0.95	E8	0.107
Inferences	= V9	=	0.71	*F2	+	0.70	E9	0.505
Reading	= F2	=	0.68	*F1	+	0.73	D2	0.464

F1 = Lexico-grammatical Knowledge
F2 = Reading Ability

* = freely estimated

Lex F = Lexical Form
Morph F = Morphosyntactic Form
Coh F = Cohesive Form
Lex M = Lexical Meaning
Coh M = Cohesive Meaning

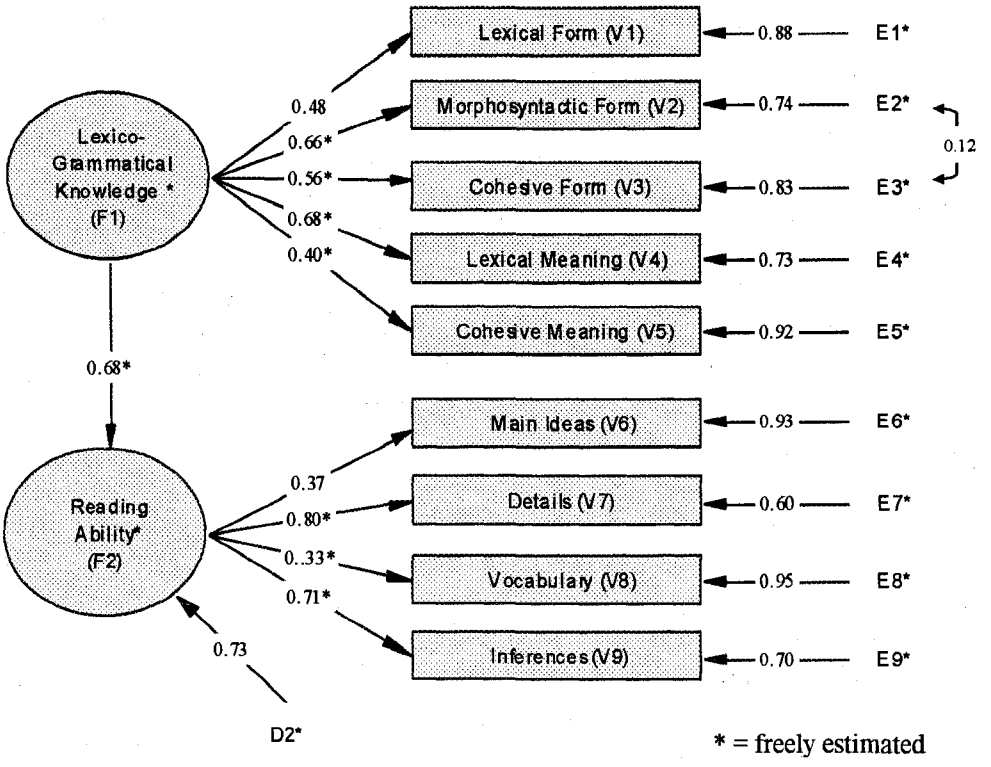
A graphical representation of the hypothesized model for the GCVR Section for Form X along with the standardized parameter estimates are presented in Figure 5.11. An inspection of Model 5.6 shows that the GCVR section for Form X is represented by a first-order factor called lexico-grammatical knowledge and reading ability measured by nine observed variables. This model supports the work of Laresen-Freeman (1991) and Purpura (2004) in their claims that lexico-grammatical knowledge consists of form and meaning. In the current study, form and meaning were represented by lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning.

This model supports the empirical study of Alderson (1993) and Purpura (1999) that lexico-grammatical knowledge is a critical linguistic resource for reading ability, and that lexico-grammatical knowledge has a strong effect on reading ability. Bernhardt (1999) reviewed the studies (e.g., Bernhardt & Kamil, 1995; Bossers, 1991; Brisbois,

1995; Carrell, 1991) which investigated the contribution of L1 reading and L2 grammar to L2 reading, and concluded that L2 reading is primarily dependent on grammatical ability in the second language. Similarly, Droop and Verhoeven (2003) found a very strong relationship between grammatical knowledge and reading ability with Dutch, Turkish, and Moroccan students in Holland. These empirical studies indicate that reading ability requires a certain threshold of lexico-grammatical knowledge in order to understand syntactic structure as well as literal and intended meaning. This was shown by the large standardized loading (0.68) of the path from lexico-grammatical knowledge to reading ability.

With regard to error terms, Model 5.6 produced one pair of correlated error terms. The correlation was 0.12, indicating that there was some redundant content being measured between morphosyntactic form and cohesive form. The redundancy is appropriate as both morphosyntactic form and cohesive form measure the form dimension of lexico-grammatical knowledge.

Figure 5.11
Model for the GCVR Section for Form X with Standardized Parameter Estimates: Model 5.6



5.3.2 Testing the Factorial Validity of the GCVR Section: Form Y

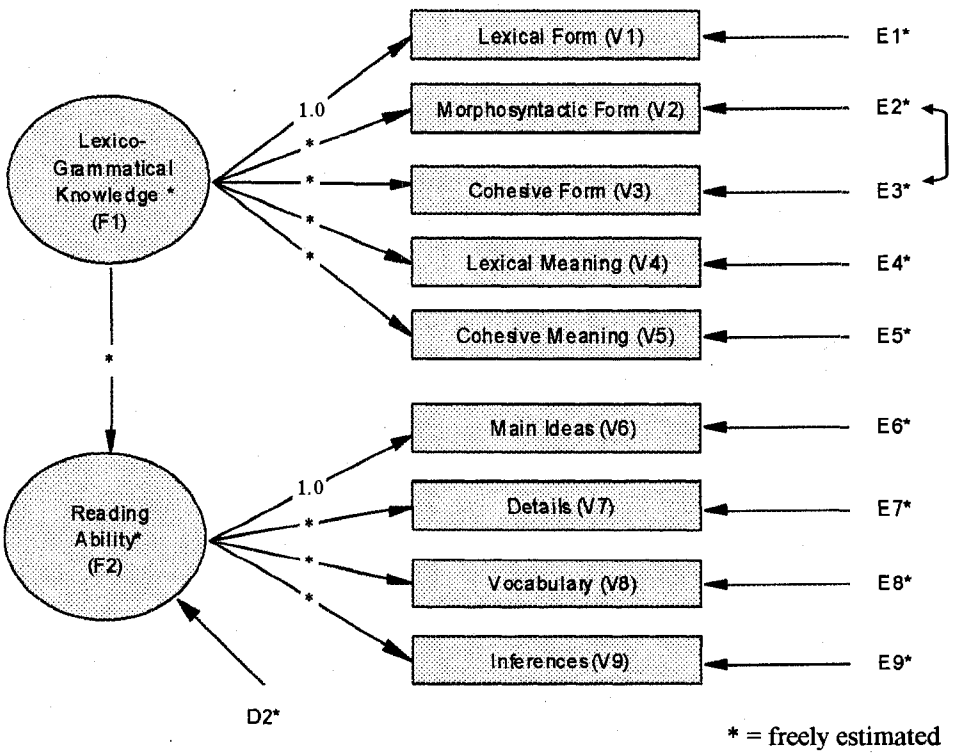
The hypothesized model: Model 5.7

Based on the substantive theory and the results of the separate CFAs for lexico-grammatical knowledge and reading ability, the GCVR section for Form Y was represented as a two-factor model of lexico-grammatical knowledge and reading ability. It contained two factors (lexico-grammatical knowledge and reading ability) with nine observed variables (lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning, denoted V1-V5 respectively, and main ideas, details, vocabulary in context, and inferences, denoted V6-9 respectively). Each observed variable was hypothesized to load on only one factor. The error terms associated with the observed

variables were postulated to be uncorrelated except the one pair of correlated error terms between morphosyntactic form and cohesive form. This pair of error term was correlated based on the results from the CFA of the GCV section for Form Y (see Model 5.3).

Based on theory and empirical research conducted by several researchers (e.g., Alderson, 1993; Bernhardt & Kamil, 1995; Bossers, 1991; Brisbois, 1995; Carrell, 1991; Droop & Verhoeven, 2003; Purpura, 1999), it was postulated that lexico-grammatical knowledge is a critical linguistic resource for reading ability, and that lexico-grammatical knowledge has a strong effect on reading ability. Therefore, the model includes a unidirectional arrow from lexico-grammatical knowledge to reading ability. This model is the same as the model for the GCVR section of Form X. This hypothesized model is presented in Figure 5.12.

Figure 5.12
Hypothesized Model for the GCVR Section for Form Y: Model 5.7



Model 5.7 is a first-order confirmatory factor analysis designed to test the multidimensionality of the GCVR section of Form Y. More specifically, it tests the hypothesis that the underlying constructs of the GCVR section is a multidimensional construct composed of lexico-grammatical knowledge and reading ability. Given the exploratory nature of the present study, the current investigation was not limited to a simple confirmation or rejection of this initial model. Rather, the relationships among the variables were explored with the goal of producing the best fitting and most substantively meaningful model. After examining the results for Model 5.7, the study proceeds to identify additional specifications and improvements if necessary.

Model 5.7 addresses the following research question:

Q2: What is the underlying trait structure of lexico-grammatical knowledge and reading ability as measured by the GCVR section in Form Y of the ECPE?

The results for Model 5.7 (Initial Model of the GCVR Section of Form Y)

Before investigating how well Model 5.7 fits the data, multivariate normality assumptions were again examined. The data produced a Mardia's coefficient of 1.01 with a normalized estimate of 6.49, suggesting multivariate non-normality. Hence, maximum likelihood robust estimation method was used to account for non-normality in the data (Bentler, 2006). Subsequently, possible multivariate outliers in the data were checked. None of the five cases listed in the EQS program were strikingly different from the others. Therefore, none of them were deleted from the analysis.

The goodness of fit index for the hypothesized two-factor model of lexico-grammatical knowledge and reading ability produced a Satorra-Bentler chi-square (S-

Bχ^2) of 821.04 with 25 degrees of freedom (df) ($p<0.01$). It also produced a CFI of 0.985 and a NNFI of 0.978. A value for the RMSEA was 0.031, which was within the acceptable range. These results indicate a good fit for the data, provided the individual parameter estimates are viable and statistically significant. Table 5.12 presents the goodness of fit summary for the hypothesized model for the GCVR section of Form Y.

Table 5.12
Results for the Hypothesized Model for
the GCVR Section for Form Y: Model 5.7

Multivariate Kurtosis	
Mardia's coefficient (G2, P)	1.01
Normalized estimate	6.49
Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	821.04
Degrees of freedom	25
Probability value for the chi-square statistic	0.0000
Fit Indices	
Comparative Fit Index (CFI)	0.985
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.978
Root Mean-Square Error of Approximation (RMSEA)	0.031

These fit statistics provided strong evidence for acceptance of Model 5.7. A series of post hoc fitting procedures (e.g., the LM and Wald tests) were performed to see whether there was a better fitting model for the sample data. However, none of the alternative models indicated by these procedures were theoretically plausible; hence, the initially-hypothesized model (Model 5.7) was treated as the best fitting model.

In examining the measurement equations of the individual parameter estimates, all estimates were found to be reasonable and statistically significant at the 0.05 level. The variances and covariances of all the independent variables were also statistically

significant at the 0.05 level. As shown in Table 5.13, the standardized solution of factor loadings for Model 5.7 were all within the acceptable limits and ranged from a low moderate 0.27 for main ideas to a high 0.81 for lexical meaning. The fairly high loading of 0.70 from grammatical knowledge to reading ability provided evidence that grammatical knowledge had an effect on reading ability in this test.

For the GCV part of the model, the factor loading patterns did not change from the original lexico-grammatical knowledge model (Model 5.3). The highest loading variables measuring lexico-grammatical knowledge were lexical meaning (0.81) and morphosyntactic form (0.70). In the current model (Model 5.7), the highest loading variables measuring grammatical knowledge were also lexical meaning (0.81) and morphosyntactic form (0.70). Not only the loading patterns were the same, but the parameter estimates were also the same between Form X and Form Y.

For the reading section, a pattern of loadings did not change from the original reading ability model (Model 5.5). The variable measuring the ability to identify details was again contributing the most to the construct of reading ability (0.73) and the variable measuring the ability to identify main ideas was again contributing the least (0.27).

With regard to standardized estimates, none of the estimates or the errors was found to be outside the acceptable range. The proportion of the variances of the measured variables accounted for by their hypothesized latent variable (i.e., R-squared) ranged from 0.073 to 0.654.

Table 5.13
Parameter Estimates for the Hypothesized Model for the GCVR Section for Form Y: Model 5.7

STANDARDIZED SOLUTION:								R-SQUARED
Lex F	=	V1	=	0.55	F1	+	0.84 E1	0.299
Morph F	=	V2	=	0.70	*F1	+	0.71 E2	0.491
Coh F	=	V3	=	0.60	*F1	+	0.80 E3	0.364
Lex M	=	V4	=	0.81	*F1	+	0.59 E4	0.654
Coh M	=	V5	=	0.38	*F1	+	0.92 E5	0.148
Main Ideas	=	V6	=	0.27	F2	+	0.96 E6	0.073
Details	=	V7	=	0.73	*F2	+	0.68 E7	0.535
Voc in Context	=	V8	=	0.30	*F2	+	0.96 E8	0.088
Inferences	=	V9	=	0.62	*F2	+	0.78 E9	0.387
Reading	=	F2	=	0.70	*F1	+	0.72 D2	0.483

F1 = Lexico-grammatical Knowledge
F2 = Reading Ability

* = freely estimated

Lex F = Lexical Form
Morph F = Morphosyntactic Form
Coh F = Cohesive Form
Lex M = Lexical Meaning
Coh M = Cohesive Meaning

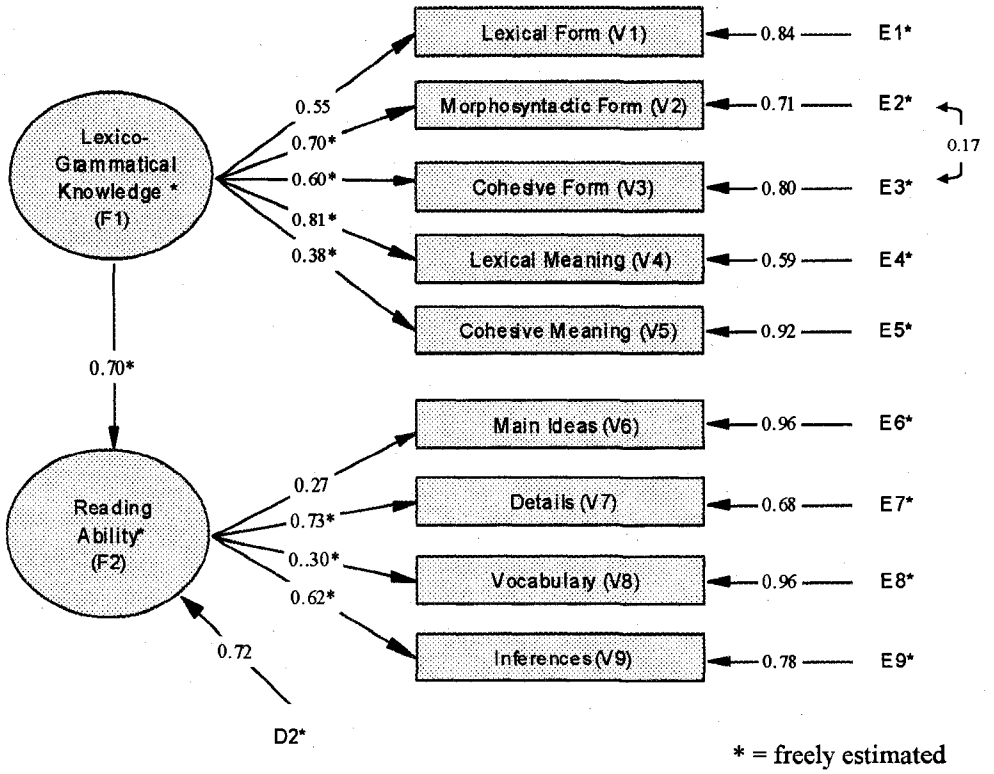
A graphical representation of the hypothesized model for the GCVR Section for Form Y along with the standardized parameter estimates are presented in Figure 5.13. An inspection of Model 5.7 shows that the GCVR section for Form Y is represented by a first-order factor called lexico-grammatical knowledge and reading ability measured by nine observed variables. This model again supports the work of Larsen-Freeman (1991) and Purpura (2004) in their claims that grammatical knowledge consists of form and meaning. In the current study, form and meaning were represented by lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning. Furthermore, this model supports the claim by Grabe (2005) that grammar resources are central to fluent reading ability.

There are empirical studies which provided evidence that grammatical knowledge has a strong effect on reading ability (Bernhardt & Kamil, 1995; Bossers, 1991; Brisbois,

1995; Carrell, 1991; Droop & Verhoeven, 2003; Purpura, 1999). These studies indicate that reading ability requires a certain threshold of lexico-grammatical knowledge in order to understand literal and intended meaning. This finding was confirmed by the large standardized loading (0.70) of the path from grammatical knowledge to reading in the current study.

With regard to error terms, Model 5.7 produced one pair of correlated error terms. The correlation was 0.17, indicating that there was some redundant content being measured between morphosyntactic form and cohesive form. The redundancy is appropriate as both morphosyntactic form and cohesive form measure the form dimension of lexico-grammatical knowledge.

Figure 5.13
Model for the GCVR Section for Form Y with Standardized Parameter Estimates: Model 5.7



5.4 Summary

This chapter investigated the fit of theoretically plausible models of the GCVR section by performing a series of separate confirmatory factor analyses. It first examined the underlying construct of the GCV section of Form X. Then, the same model was tested for Form Y. The results indicated that the GCV section for Form X and Form Y are represented by a first-order factor called lexico-grammatical knowledge measured by five observed variables (lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning). The model consisted of one pair of error terms, which was between morphosyntactic form and cohesive form.

Once the model for the GCV section was confirmed, the hypothesized model for the reading section was examined for Form X. Then, the same model was tested for Form Y. The results indicated that the reading section for Form X and Form Y are represented by a first-order factor called reading ability measured by five observed variables (the ability to identify main ideas, the ability to identify details, the meaning of vocabulary in context, and the ability to identify inferences).

After the GCV and reading sections were modeled separately, they were combined to examine the underlying trait structure of the GCVR section. It first examined the underlying construct of the GCVR section of Form X. Then, the same model was tested for Form Y. The results indicated that the GCVR section for Form X and Form Y are represented by a first-order factor called lexico-grammatical knowledge and reading ability measured by nine observed variables (lexical form, morphosyntactic form, cohesive form, lexical meaning, cohesive meaning, the ability to identify main ideas, the ability to identify details, the meaning of vocabulary in context, and the ability to identify inferences). The model consisted of one pair of error terms, which was between morphosyntactic form and cohesive form.

In the next chapter simultaneous multi-group analyses are performed on the models of the GCVR section to examine the degree to which the models are equivalent across Form X and Form Y.

Chapter VI

MULTI-GROUP ANALYSES

The establishment of separate baseline models of the underlying trait structures of the GCVR section for Form X and Form Y is a requirement for testing hypotheses related to cross-group (i.e., form) invariance (Byrne, 2006). As a result, two separate models of the underlying trait structures of the GCVR section were established in the previous chapter. Although the underlying trait structures of the GCVR sections were the same, this does not necessarily indicate that their measurement and structural models exhibit equivalence across forms when estimated simultaneously (Purpura, 1999). Estimating models simultaneously is a much more stringent hypothesis. In this chapter, simultaneous multi-group analyses are performed to examine the degree to which the models are equivalent across Form X and Form Y.

6.1 Testing for Configural Invariance

The initial step in testing for invariance is to examine the configural invariance (Byrne, 2006). This means that the same factor structure must hold across forms. At this point, no equality constraints are imposed on the parameters. That is, the parameters estimated in the baseline model for each form separately are again estimated in this multi-group model (Byrne, 2006). In other words, the model tested here is a multi-group representation of the baseline models. It includes the baseline models for Form X and Form Y within the same file. There are two reasons why configural invariance is tested.

First, it allows for invariance tests to be performed simultaneously across two forms. Second, in testing for invariance, the fit of this configural model gives the baseline value against which subsequently specified invariance models are compared (Byrne, 2006). Unlike the single-group analyses, this test provides only one set of fit statistics for overall model fit.

Prior to examining the goodness of fit summary, multivariate normality assumptions were checked. Because the models of the GCVR section for both Form X and Form Y were multivariate non-normal, the estimation for the configural model was based on the robust statistics. Results yielded from the testing of this configural model are shown in Table 6.1.

Table 6.1
Results for Testing Configural Invariance

Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	1494.48
Degrees of freedom	50
Probability value for the chi-square statistic	0.00000
Fit Indices	
Comparative Fit Index (CFI)	0.985
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.978
Root Mean-Square Error of Approximation (RMSEA)	0.030

Goodness of fit statistics related to this model produced a fitting multi-group model, with the S-Bχ^2 of 1494.48 with 50 degrees of freedom. It produced a CFI of 0.985 and a NNFI of 0.978. A value for the RMSEA was within the acceptable limit. Overall, the goodness of fit indices suggest that the configural model fit the data well. The results

indicate that the structure of the GCVR section is represented as a first-order model, with the pattern of factor loadings specified in accordance with this initial multi-group model.

6.2 Testing for Measurement Invariance

The next step in testing for invariance was to determine equality with respect to the measurement model. In this test, the invariance of factor loadings was of interest. All the paths were constrained except the ones which were fixed to 1.0 for identification purposes.

The errors for the observed variables were not constrained to be equal, as they were generally uncorrelated with other variables (Purpura, 1999), and the primary concern of the current study was to investigate the invariance of factor loadings, the path coefficients, and the factor invariances. Moreover, tests for the measurement error variances-covariances are not generally examined as it is considered excessively stringent (Byrne, 2006). Hence, the current study focused on the invariance of factor loadings, and the errors for the observed variables were not constrained to be equal.

There were seven parameters constrained to equal for the test for measurement invariance: (1) lexico-grammatical knowledge to morphosyntactic form, (2) lexico-grammatical knowledge to cohesive form, (3) lexico-grammatical knowledge to lexical meaning, (4) lexico-grammatical knowledge to cohesive meaning, (5) reading ability to details, (6) reading ability to vocabulary in context, (7) reading ability to inferences.

Results from the testing of the measurement invariance are shown in Table 6.2. Conventionally, it is asserted that invariance holds if goodness-of-fit related to the model testing for measurement invariance is considered adequate (Widaman & Reise, 1997),

and if there is minimal difference in fit from that of the configural model (Byrne, 2006). The results showed that the multi-group model with the equality constraints on seven parameters underwent some deterioration in model fit (corrected $\Delta S\text{-}B\chi^2 = 845.82$, $p < 0.01$; $\Delta^*CFI = 0.006$) from the configural model. Overall, however, the multi-group model still exhibited a good fit to the data. It produced a CFI of 0.979, a NNFI of 0.973, and a value for the RMSEA was 0.036.

Table 6.2
Results for Testing Measurement Invariance

Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	2340.77
Degrees of freedom	57
Probability value for the chi-square statistic	0.00000
Fit Indices	
Comparative Fit Index (CFI)	0.979
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.973
Root Mean-Square Error of Approximation (RMSEA)	0.036

In order to determine which parameters were found to be invariant across forms, the Lagrange Multiplier test was examined. If the parameters are invariant across forms, the probability values associated with the incremental univariate χ^2 values should be >0.05 (Byrne, 2006). Review of these values, as reported in Table 6.3, revealed that none of the probability values of the constrained parameters were greater than 0.05. In fact, the probability values for all seven parameters were 0.000. In other words, according to the LM test, none of the constrained parameters operated equivalently across forms. This suggests that the underlying theoretical models estimated for Form X and Form Y were not equivalent, in as much as they have statistically different parameter values.

Table 6.3
Testing Invariance of Measurement Model: LM Test Statistics

	Cumulative Multivariate Statistics			Univariate Increment	
Parameter	χ^2	D.F.	Probability	χ^2	Probability
(Form X, V8, F2) – (Form Y, V8, F2) = 0	200.78	1	0.000	200.78	0.000
(Form X, V4, F1) – (Form Y, V4, F1) = 0	385.43	2	0.000	184.65	0.000
(Form X, V3, F1) – (Form Y, V3, F1) = 0	486.59	3	0.000	101.16	0.000
(Form X, V2, F1) – (Form Y, V2, F1) = 0	684.37	4	0.000	197.78	0.000
(Form X, V5, F1) – (Form Y, V5, F1) = 0	814.93	5	0.000	130.57	0.000
(Form X, V7, F2) – (Form Y, V7, F2) = 0	846.39	6	0.000	31.46	0.000
(Form X, V9, F2) – (Form Y, V9, F2) = 0	890.74	7	0.000	44.35	0.000

V2 = Morphosyntactic Form
V3 = Cohesive Form
V4 = Lexical Meaning
V5 = Cohesive Meaning
F1 = Lexico-grammatical Knowledge
V7 = Details
V8 = Vocabulary in Context
V9 = Inferences
F2 = Reading Ability

Although the differences in the constrained parameters turned out significant in the LM test, the actual values of the parameters were very similar across forms. The parameter estimates were compared across forms when the model was measured simultaneously (see Table 6.4). All parameter estimates were closely examined to see how different the values were.

Out of ten parameters simultaneously estimated in the model, nine of them were within 0.1 difference. The one with greater than 0.1 difference was cohesive form. The parameter estimates for cohesive form were 0.50 for Form X and 0.64 for Form Y. This indicates that lexico-grammatical knowledge was better explained by cohesive form in Form Y than in Form X.

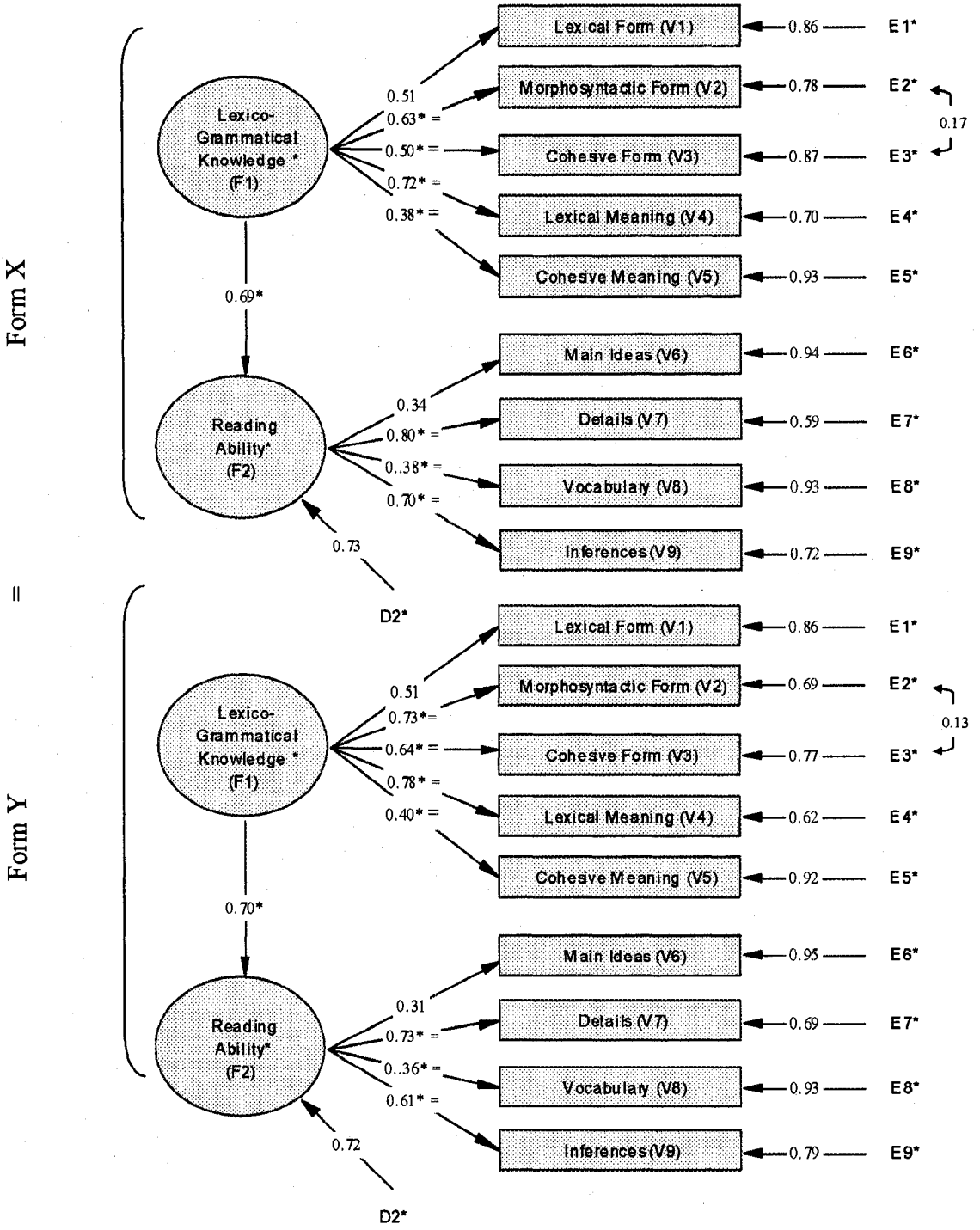
Table 6.4
Simultaneous Group Analysis with the Constrained Parameters in Measurement Model

			β_X	β_Y			ε_X	ε_Y	
Lex F	= V1	=	0.51	0.51	F1	+	0.86	0.86	E1
Mor F	= V2	=	0.63	0.73	*F1	+	0.78	0.69	E2
Coh F	= V3	=	0.50	0.64	*F1	+	0.87	0.77	E3
Lex M	= V4	=	0.72	0.78	*F1	+	0.70	0.62	E4
Coh M	= V5	=	0.38	0.40	*F1	+	0.93	0.92	E5
Main	= V6	=	0.34	0.31	F2	+	0.94	0.95	E6
Det	= V7	=	0.80	0.73	*F2	+	0.59	0.69	E7
Voc	= V8	=	0.38	0.36	*F2	+	0.93	0.93	E8
Infer	= V9	=	0.70	0.61	*F2	+	0.72	0.79	E9
Read	= F2	=	0.69	0.70	*F1	+	0.73	0.72	D2

Lex F = Lexical Form
Mor F = Morphosyntactic Form
Coh F = Cohesive Form
Lex M = Lexical Meaning
Coh M = Cohesive Meaning
Main = Main Ideas
Det = Details
Voc = Vocabulary in Context
Infer = Inferences
F1 = Lexico-grammatical Knowledge
F2 = Reading

The numerical estimates of the parameter values were not identical across forms; however, the differences in many of the parameters appear to be marginal and the parameter estimates were substantively or "practically" equivalent. Considering the large sample size of this data, even a tiny numerical difference in the parameter estimates can result in statistically significant difference. Based on these results, it cannot be concluded that these two forms are equivalent when they are simultaneously modeled. However, it may be reasonable to suggest that the two forms have the identical underlying trait structures despite the marginal differences in the parameter estimates. The diagrammatic representation of the parameter estimates are shown in Figure 6.1.

Figure 6.1
Simultaneous Group Analysis of the Constrained Parameters in the Measurement Model: Form X and Form Y



* = freely estimated

6.3 Testing for Structural Invariance

Following the testing for configural invariance and measurement invariance, the final step in testing for invariance was to address equality with respect to the structural model. In this test, the models for Form X and Form Y were estimated simultaneously with all the factor loadings in the measurement models, the path coefficients in the structural models, and the factor variances constrained to be equal across forms. This is the most constrained model among the three models discussed so far (i.e., configural model, measurement model, and structural model).

When estimated simultaneously, the fully-constrained, multi-group model of lexico-grammatical knowledge and reading ability produced a CFI of 0.976, a NNFI of 0.970, and RMSEA of 0.038, suggesting a good model data fit. Furthermore, all parameter estimates were significant at the 0.05 level (see Table 6.5). Again, the results showed that the multi-group model underwent some deterioration in model fit (corrected $\Delta S\text{-}\ B\chi^2 = 1149.03$, $p < 0.01$; $\Delta^*CFI = 0.009$) from the configural model. Overall, however, the multi-group model still exhibited a good fit to the data.

Table 6.5
Results for Testing Structural Invariance

Goodness of Fit Summary Method = Robust	
Satorra-Bentler scaled chi-square	2638.51
Degrees of freedom	58
Probability value for the chi-square statistic	0.00000
Fit Indices	
Comparative Fit Index (CFI)	0.976
Bentler-Bonnet Non-Normed Fit Index (NNFI)	0.970
Root Mean-Square Error of Approximation (RMSEA)	0.038

Just as in testing for measurement invariance, it was necessary to determine which parameters were found to be invariant across forms using the LM test. If the same parameters are invariant across forms, the probability values associated with the incremental univariate χ^2 values should be >0.05 (Byrne, 2006). Review of these values, as reported in Table 6.6, revealed that none of the probability values of the constrained parameters were greater than 0.05. In other words, according to the LM test, none of the parameters operated equivalently across forms. This suggests that Form X and Form Y were not considered equivalent when the parameters were constrained in both measurement model and structural model.

Table 6.6
Testing Invariance of Structural Model: LM Test Statistics

	Cumulative Multivariate Statistics			Univariate Increment	
Parameter	χ^2	D.F.	Probability	χ^2	Probability
(Form X, V4, F1) – (Form Y, V4, F1) = 0	339.40	1	0.000	339.40	0.000
(Form X, V8, F2) – (Form Y, V8, F2) = 0	519.80	2	0.000	180.41	0.000
(Form X, F2, F1) – (Form Y, F2, F1) = 0	696.98	3	0.000	177.78	0.000
(Form X, V3, F1) – (Form Y, V3, F1) = 0	805.75	4	0.000	108.77	0.000
(Form X, V2, F1) – (Form Y, V2, F1) = 0	991.70	5	0.000	185.95	0.000
(Form X, V5, F1) – (Form Y, V5, F1) = 0	1090.75	6	0.000	99.05	0.000
(Form X, V7, F2) – (Form Y, V7, F2) = 0	1117.74	7	0.000	26.99	0.000
(Form X, V9, F2) – (Form Y, V9, F2) = 0	1158.28	8	0.000	40.54	0.000

V2 = Morphosyntactic Form
V3 = Cohesive Form
V4 = Lexical Meaning
V5 = Cohesive Meaning
F1 = Lexico-grammatical Knowledge
V7 = Details
V8 = Vocabulary in Context
V9 = Inferences
F2 = Reading Ability

Although the differences in all eight constrained parameters were statistically significant in the LM test, the actual values of the parameters were very similar across forms. The parameter estimates for both forms are presented in Table 6.7. All parameter estimates were closely examined to see how different the values were.

Of the ten parameters simultaneously estimated in the model, eight were within 0.1 difference. One of the variables with a difference greater than 0.1 was cohesive form, the other was reading ability.

The cohesive form variable had noticeable differences in the test for measurement model. The parameter estimates for cohesive form were 0.51 for Form X and 0.63 for Form Y. This indicates that lexico-grammatical knowledge was better explained by cohesive form in Form Y than in Form X.

The estimates for the parameters connecting lexico-grammatical knowledge to reading ability in the fully-constrained structural model were different from the

corresponding unconstrained parameter estimates in the measurement model. In other words, when those parameters were not constrained in the measurement model, the estimates were 0.69 for Form X and 0.70 for Form Y. When they were constrained in the structural model, the estimates were 0.63 for Form X and 0.74 for Form Y. This indicates that the parameter between grammatical knowledge and reading ability was not invariant across forms. It shows that the degree of effect from lexico-grammatical knowledge to reading ability was different across forms; however, the values 0.63 and 0.74 indicate that there was a strong effect from lexical-grammatical knowledge to reading ability in both forms.

Table 6.7
Simultaneous Group Analysis with the Constrained Parameters in Structural Model

			β_X	β_Y			ε_X	ε_Y	
Lex F	= V1	=	0.52	0.50	F1	+	0.85	0.87	E1
Mor F	= V2	=	0.65	0.72	*F1	+	0.76	0.70	E2
Coh F	= V3	=	0.51	0.63	*F1	+	0.86	0.77	E3
Lex M	= V4	=	0.73	0.77	*F1	+	0.69	0.63	E4
Coh M	= V5	=	0.39	0.40	*F1	+	0.92	0.92	E5
Main	= V6	=	0.33	0.32	F2	+	0.95	0.95	E6
Det	= V7	=	0.80	0.74	*F2	+	0.61	0.68	E7
Voc	= V8	=	0.37	0.37	*F2	+	0.93	0.93	E8
Infer	= V9	=	0.69	0.63	*F2	+	0.73	0.78	E9
Read	= F2	=	0.63	0.74	*F1	+	0.78	0.68	D2

Lex F = Lexical Form
Mor F = Morphosyntactic Form
Coh F = Cohesive Form
Lex M = Lexical Meaning
Coh M = Cohesive Meaning
Main = Main Ideas
Det = Details
Voc = Vocabulary in Context
Infer = Inferences
F1 = Lexico-grammatical Knowledge
F2 = Reading

The numerical estimates of the parameter values were not identical across forms; however, the differences in many of the parameters appear to be marginal and the

parameter estimates are substantively equivalent. Considering the large sample size of this data, even a tiny numerical difference in the parameter estimates can result in statistically significant difference.

There are some possible explanations for the differences in the parameter estimates. First, the population was assumed to be identical, but they may be somewhat different. For example, when the central tendency of the age distribution was examined, Form X and Form Y appeared basically equivalent across forms. The mean age was 21.10 for Form X and 21.92 for Form Y and the standard deviation was 6.73 and 6.63 respectively. The median age was 20 for both forms. However, despite these very similar mean estimates, the large sample sizes (N=33662 for Form X and N=32473 for Form Y) actually imply a statistically significant difference in the population means.

To illustrate this point, a 95 percent confidence interval for the mean age in the population as a whole was calculated for both Form X and Form Y. The formula for calculating a confidence interval of 95 percent (assuming the underlying distributions of the population data are independent and normal) is presented below (Gujarati, 1988):

$$\overline{X} - 1.96\left(\frac{\sigma}{\sqrt{N}}\right) \leq \mu \leq \overline{X} + 1.96\left(\frac{\sigma}{\sqrt{N}}\right)$$

where σ is the standard deviation, $\overline{X}$ is the mean, and N is the population.

By substituting the estimated values of $\overline{X}$, σ, and N, the 95 percent confidence interval for mean age of those who took Form X is: $21.03 \leq \mu \leq 21.17$. For Form Y, the 95 percent confidence interval is: $21.85 \leq \mu \leq 21.99$. The age differences in the mean age of the populations taking Form X (21.10) and Form Y (21.92) are relatively small in value (less than a year), but the values at the significance level of 0.05 do not overlap with each other. In other words, the difference in the population means is estimated to be

statistically significant, even though the differences in the mean is only 0.82. This is due to the large sample size. More formally, a t-test of zero difference in means was conducted, that test produced a T-statistic of -15.78 which is statistically significant at any conventional level:

$$t = \frac{\bar{X}_X - \bar{X}_Y}{\sqrt{\frac{s_X^2}{n_X - 1} + \frac{s_Y^2}{n_Y - 1}}}$$

$$= \frac{21.10 - 21.92}{\sqrt{\frac{6.73^2}{33662 - 1} + \frac{6.63^2}{32473 - 1}}}$$

$$= -15.78$$

In addition to the difference in means, other aspects of the age distributions of the test-takers appeared somewhat different. For instance, 30.62 percent of the test-takers were age 13-16 for Form X, whereas only 22.81 percent of the test-takers were in that age range for Form Y (See Table 3.2 on P. 70). Generally, Form Y consisted of relatively older test-takers than Form X.

This discussion of differences in the mean age of test-takers is not intended as an argument that age differences in the population were the only factor leading to the finding of significant differences in the multi-group model across forms. Instead, the discussion is intended, first, to demonstrate that the large sample size can render even relatively small differences in population statistics (means) statistically significant, and, second, to provide an example of at least one potential explanation for the finding of non-invariance in the multi-group model.

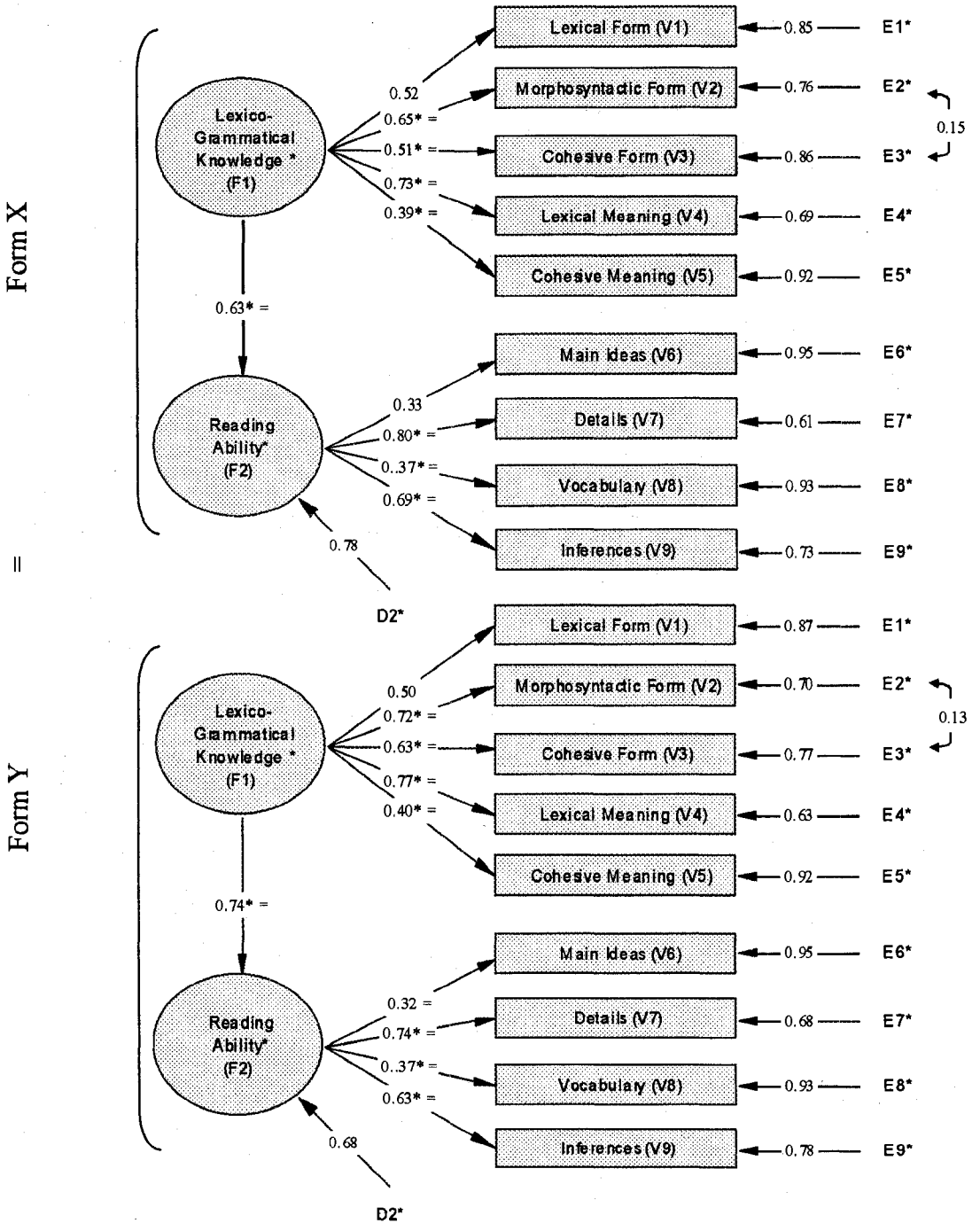
Another reason for the differences may be due to other characteristics of the participants. Other than the native language background and the age distribution data, none of the information on the participants was provided. There may have been differences in other characteristics such as educational background, socio-economic status, which may have caused differences in the analyses.

Another reason for the differences may be grammatical features tested in the GCV section. As shown in Table 3.7 (P. 85), the grammatical features tested in the GCV section were somewhat different in Form X and Form Y. Although the number of items measuring the variables (morphosyntactic form, lexical form, etc.) were the same, more detailed analysis shows that the grammatical features tested were slightly different.

Another possible reason for the differences may be the coding accuracy. Although the items were coded by experienced coders, their codings may still not have accurately or consistently categorized what is measured by each item. As Alderson (1990) claims, it is frequently difficult to get expert judges to agree on what ability is operationalized by which test item.

In summary, while the results indicated that none of the parameters were invariant across forms, the differences in the parameter estimates were marginal. Moreover, the two forms had the identical underlying trait structures. Therefore, it may be reasonable to suggest that the two forms were comparable. The diagrammatic representation of the parameter estimates are shown in Figure 6.2.

Figure 6.2
Simultaneous Group Analysis of the Constrained Parameters in the Structural Model: Form X and Form Y



6.4 Criteria in Determining Evidence of Invariance

In reporting on evidence of invariance, it is a common practice to report the difference in χ^2 values derived from the comparison of χ^2 values associated with numerous models under test with the baseline configural model (Byrne, 2006). Yuan and Bentler (2004) suggested that evidence in support of multigroup invariance should be based on the $\Delta\chi^2$ test for every SEM application. The current study used Satorra-Bentler scaled chi-square; hence, a correction to the value was needed.[7] This is because the difference in S-Bχ^2 is not distributed as χ^2 (Bentler, 2006). If the difference value is statistically significant, it indicates that the constraints specified in the more restrictive model do not hold (Byrne, 2006). In other words, the two models are not considered equivalent across forms. If, on the other hand, the difference in chi-square value is statistically non-significant, it indicates that all specified equality constraints are acceptable.

The corrected ΔS-Bχ^2 for the current study was calculated. From the configural model to the measurement model, the corrected ΔS-Bχ^2 was 845.82. This was statistically significant, which indicated that the constraints specified in the measurement model did not hold. In other words, the two models of grammatical knowledge and reading ability for Form X and Form Y did not appear to be equivalent.

From the configural model to the structural model, the corrected ΔS-Bχ^2 was 1149.03 with the statistical significance. This suggested that the constraints specified in the structural model again did not hold. In other words, the two models produced for Form X and Form Y were also not considered equivalent.

---

[7] The correction formula for computing ΔS-Bχ^2 value is provided in Byrne (2006).

The evaluative strategy using the $\Delta\chi^2$ or ΔS-Bχ^2 in determining evidence of measurement invariance (Byrne, 2006) is considered a conventional approach, which was based on Jöreskog's (1971) method in testing for multi-group equivalence. While this method might be a useful criterion in some contexts, researchers have argued that the $\Delta\chi^2$ value is "as sensitive to sample size and non-normality as the χ^2 statistic itself; thereby rendering it an impractical and unrealistic criterion on which to base evidence of invariance" (Byrne, 2006, p. 247). As a result, there is a tendency to argue for evidence of invariance based on other criteria.

An alternative criteria relates to whether the multi-group model exhibits an adequate fit to the data. The second focuses on the ΔCFI values between the models. There are several theories on what should be an appropriate cut-off value of the ΔCFI. For example, Little (1997) suggested that the difference should not exceed 0.05, whereas Cheung and Rensvold (2002) suggested the value be less than 0.01. The current study used the cut-off value of 0.01 as it is more conservative than 0.05.

Table 6.8 contains results from the tests for invariance of the model for the GCVR section across forms. Presented in this table are goodness of fit statistics related to all three models, along with the ΔS-Bχ^2 values, Δdf values, and ΔCFI values from their comparisons with the configural model. The configural model was used as the baseline against the other two models in order to determine evidence of invariance. Review of these results determines the extent of invariance based on the ΔCFI values.

For the comparison between the configural model and the measurement model, the ΔCFI was only 0.006, which was within the acceptable limit (Cheung & Rensvold, 2002). This suggested that the models were comparable across forms when the

parameters were constrained between the observed variables and their factors. The comparison between the configural model and the structural model revealed that the ΔCFI was 0.009, which was again within the acceptable limit. This again suggested that the models were comparable across forms when all the parameters in the measurement models and the path coefficient in the structural models were constrained and simultaneously estimated.

In summary, the examination of the significance of the corrected ΔS-Bχ^2 values indicated that none of the models were equivalent across forms. On the other hand, the examination of ΔCFI values suggested that the constrained parameters in the measurement model were invariant across forms. In other words, Form X and Form Y were comparable when the parameters were constrained both in the measurement model and the structural model.

As shown here, different evaluative criteria (i.e., significance of the corrected ΔS-Bχ^2 values or the ΔCFI values) revealed different outcomes of the multi-group invariance. In other words, researchers can draw a different conclusion from the same models depending on whether they look at the significance of the corrected ΔS-Bχ^2 values or the ΔCFI values to determine the multi-group invariance. Therefore, it is important to report both values to avoid drawing inaccurate conclusions.

Table 6.8
Tests for Invariance across Forms: Summary of Goodness of Fit Statistics

Model	S- Bχ	df	*CFI	*NNFI	*RMSEA	Corrected ΔS- Bχ²	Δ df	Δ*CFI
Model 1 Configural Model No Constraints	1494.18	50	0.985	0.978	0.030	-----	----	-----
Model 2 Measurement Model Invariant (factor loadings)	2340.77	57	0.979	0.973	0.036	845.82	7	0.006
Model 3 Structural Model Invariant (factor covariances)	2638.51	58	0.976	0.970	0.038	1149.03	8	0.009

6.5 Summary

In this chapter, simultaneous multi-group analyses were performed to examine the degree to which the models are equivalent across Form X and Form Y. In order to do that, the models were examined in three steps: testing for configural invariance, measurement invariance, and structural invariance. When the models were estimated, a number of equality constraints were imposed on the parameters.

The results showed that the tests of invariance across forms could not be uniformly supported in the data. When the equality constraints were imposed on the parameters, the differences in the parameters were found to be statistically significant. In other words, the parameters were statistically different across forms. However, when the values of the parameter estimates were examined, the estimates were similar in value across forms. The differences in the estimated parameter values appeared to be quantitatively marginal, and it appeared reasonable to suggest that the parameter

estimates were substantively equivalent, despite a statistically significant difference in the numerical estimates of the parameter values.

When the extent of invariance was examined using the corrected $\Delta S\text{-}B\chi^2$, the results suggested that neither the measurement model nor the fully-constrained structural model were equivalent across forms. However, when the extent of invariance was reviewed using the ΔCFI values, the results suggested that the models were comparable across forms in both measurement model and fully-constrained structural model. Given these contradictory results it may be reasonable to suggest that Form X and Form Y were not proven to be strictly statistically equivalent across forms. However, despite the issue of technically different parameter values, these parameter estimates were relatively close together with similar signs, and of similar magnitude. Moreover, the goodness-of-fit criteria has shown that the models fit the data well when simultaneously estimated.

Chapter VII

CONCLUSIONS

The primary purpose of this study was to investigate the comparability of the underlying trait structure of the two test forms of the grammar/cloze/vocabulary/reading (GCVR) section of the Examination for the Certificate of Proficiency in English (ECPE).

In this concluding chapter, the results of the present study are summarized as they relate to the research questions posed in Chapter 1. Then, theoretical, methodological, and pedagogical implications of the study are discussed. Finally, some suggestions for further research are presented.

7.1 Summary of the Results

This section provides a brief summary of the research results in the form of answers to the three research questions posed in Chapter 1.

Research Question 1: *What is the underlying trait structure of lexico-grammatical knowledge and reading ability as measured by the GCVR section in Form X of the ECPE?*

In response to this question, a series of confirmatory factor analyses (CFAs) was performed, in which hypotheses related to the nature of lexico-grammatical knowledge and reading ability were posited and tested. In the end, a model with nine measured variables and one pair of correlated errors provided a reasonable explanation of the relationship among the observed variables. Lexico-grammatical knowledge is measured

by five observed variables called lexical form, morphosyntactic form, cohesive form, lexical meaning, and cohesive meaning. Reading ability is measured by four observed variables called the ability to identify main ideas, the ability to identify details, the meaning of vocabulary in context, and the ability to make inferences. There was one pair of variables with correlated errors: morphosyntactic form and cohesive form.

The fairly high loading estimate of 0.68 from lexico-grammatical knowledge to reading ability provided strong evidence that lexico-grammatical knowledge is a key predictor of reading ability in this test. This model is supported by many empirical studies (e.g., Alderson, 1993; Bernhardt & Kamil, 1995; Bossers, 1991; Brisbois, 1995; Carrell, 1991; Purpura, 1999) that lexico-grammatical knowledge is a critical linguistic resource for reading ability, and that lexico-grammatical knowledge has a strong effect on reading ability. More recently, Droop and Verhoeven (2003) found a very strong relationship between grammatical knowledge and reading ability with Dutch, Turkish, and Moroccan students in Holland. These empirical studies along with the results from the current study indicate that reading ability requires a certain threshold of lexico-grammatical knowledge in order to understand syntactic structure as well as literal and intended meaning.

Research Question 2: *What is the underlying trait structure of lexico-grammatical knowledge and reading ability as measured by the GCVR section in Form Y of the ECPE?*

In response to research question 2, a series of CFAs were again performed to examine the underlying trait structure of lexico-grammatical knowledge and reading ability. The data supported a two-factor model of grammatical knowledge and reading

ability with nine measured variables and three pairs of correlated errors. This model was the same model as the one for Form X.

The factor loading from lexico-grammatical knowledge to reading ability was estimated to be 0.70, suggesting that lexico-grammatical knowledge is a critical linguistic resource for reading ability, and that lexico-grammatical knowledge has relatively strong effect on reading ability.

Research Question 3: *To what extent does the GCVR section measure the same underlying trait structure across the different ECPE test forms?*

In order to examine the degree to which the models are equivalent across Form X and Form Y, a series of simultaneous multi-group analyses were performed. The models of Form X and Form Y were examined in three steps: testing for configural invariance, measurement invariance, and structural invariance. When the models were estimated, a number of equality constraints were imposed on the parameters.

The results indicated that the tests of invariance across forms could not be uniformly supported in the data. When the equality constraints were imposed on the parameters, the differences in the parameters were found to be statistically significant. In other words, the constrained parameter values were different across forms. However, when the values of the parameter estimates were examined, the numbers were quite similar in value across forms. The differences in the parameter values appeared to be quantitatively marginal, and it appeared reasonable to suggest that the parameter estimates were substantively equivalent, despite a statistically significant difference in the numerical estimates of the parameter values.

7.2 Implications of the Study

In this section, theoretical, methodological, and practical implications of the study are discussed.

7.2.1 Theoretical Implications

The present study has extended the generalizability and representativeness of Purpra's (2004) models of grammatical knowledge. It was able to provide supporting evidence that his model could work well for a test measuring test-taker's grammatical knowledge. Chang (2004) was the first to examine the fit of Purpura's model, but his study only focused on the nature of L2 grammatical knowledge as it relates to relative clause. The current study, on the other hand, used data from a standardized test which assesses a variety of grammatical features. The results of the present study provided valuable information that the GCV section of the ECPE supports Purpura's model of grammatical knowledge, which is composed of form and meaning.

7.2.2 Methodological Implications

From a methodological point of view, this study has demonstrated the significance of using structural equation modeling (SEM). SEM has presented evidence that it can be a powerful research tool for investigating the underlying trait structures of latent factors and for providing insights into the interrelationships among the latent factors as well as the observed variables.

Furthermore, this study used SEM to measure comparability of the underlying trait structures across different test forms. It is a common practice to use the simultaneous multi-group SEM to compare different groups (e.g., Bae & Bachman, 1998; Ginther & Stevens, 1998; Kunnan, 1995; Purpura, 1998, 1999; Pyo, 2001; Yun, 2005), but this study is the first in the field of applied linguistics to use the simultaneous multi-group SEM to compare different test forms. This may provide valuable insights to applied linguists that the use of multi-group SEM in assessing comparability of test forms may be a new focus on complementing the IRT test equating procedure.

7.2.3 Practical Implications

When the underlying trait structure was separately modeled, Form X and Form Y of the GCVR section produced identical factorial models. In other words, the same model fit the data well for each form. Moreover, the parameter estimates were very similar in Form X and Form Y, which suggested that the two forms are comparable when modeled separately.

When the underlying trait structure of each model was estimated simultaneously with all the factor loadings in the measurement models and the path coefficient in the structural model constrained to be equal across forms, the results showed that the invariance across forms was not supported in the data. In other words, when the equality constraints were imposed on the parameters, the differences in the parameters were found to be statistically significant. This suggests that the two forms were not identical when modeled simultaneously.

There are several possible reasons to explain these results. First of all, while the population appeared slightly different in Form X and Form Y. More specifically, the participants who took Form Y were somewhat older than the participants in Form X. The results for the multi-group analyses may have been different if the populations were identical.

Second, differences in the tested grammatical features in Form X and Form Y may have caused the constrained parameter estimates to be statistically significant. Although the number of items measuring the variables (e.g., morphosyntactic form, lexical form, etc.) were the same, more detail analyses shows that the grammatical features tested were slightly different. For example, the ability to understand formulaic expressions was tested in Form Y, but not in Form X. The ability to recognize pronouns and reference was tested in Form X, but not in Form Y. If the grammatical features tested in Form X and Form Y were more controlled, the results of the multi-group analyses may have provided evidence that these two test forms were invariant.

Another possible reason for the differences may be coding accuracy. The items were coded by experienced coders; however, their codings may still not have accurately or consistently categorized what is measured by each item. This may have caused the constrained parameter estimates to be variant across forms.

Although there are many possible explanations for the fact that the invariance across forms was not supported by the data, the test developers may find the results useful in revising the test items for the future. It may be helpful if they could better control the grammatical features tested in each form. For example, if formulaic expression is tested in Form X, it should also be tested in Form Y.

7.3 Suggestions for Further Research

This study used a sophisticated statistical tool to investigate the comparability across two test forms. Although the results of this study are beneficial, it could profit from further in-depth investigation.

One area of future research relates to the models used in this study. The two-factor model of lexico-grammatical knowledge and reading ability was selected as a model that would best represent the underlying trait structures of the GCVR section of ECPE for both Form X and Form Y; however, there may be other models that would be a better representative of the GCVR section. It would be interesting to see if other theoretical models fit the data better than the model used in the current study.

A second area of future research identified by this study is an investigation of multiple test forms. The current study only compared the underlying trait structures of two different test forms. The results may have been different if more forms of the test were analyzed. Hence, it would be interesting to see if the underlying trait structures of three or more test forms are simultaneously compared with one another.

A third area of future research relates to coding accuracy. The current study used expert judges in coding test items; however, the inter-coder agreement was not as high as it could have been. This may be caused by the insufficient explanation in operationalizing the variables in the current study. If there were a higher inter-coder agreement in the data, the items may have been better categorized, reflecting more accurately the nature of the test items.

Although there are many limitations, the findings of this study have contributed to a deeper understanding of the underlying trait structures of the GCVR section of ECPE as well as the comparability of the construct across test forms. It would be interesting to see if the multi-group SEM comparison approach can be used for tests other than ECPE.

REFERENCES

Adams, M. (1990). *Beginning to read.* Cambridge, MA: MIT Press.

Aebersold, J., & Field, M. (1997). *From reader to reading teacher.* Cambridge: Cambridge University Press.

Alderson, J. C. (1990). Testing reading comprehension skills (Part One). *Reading in a Foreign Language 6*, 425-438.

Alderson, J. C. (1991). Language testing in the 1990s: How far have we come? How much further have we to go? In S. Anivan (Ed.), *Current developments in language testing* (pp. 1-26). Singapore: SEAMEO Regional Language Centre.

Alderson, J. C. (1993). The relationship between grammar and reading in an English for academic purposes test battery. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium* (pp. 203-219). Alexandria, VA: TESOL.

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Allison, D. (1999). *Language testing and evaluation: An introductory course.* Singapore: Singapore University and World Scientific.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anderson, N. (1999). *Exploring second language reading.* Boston: Heinle & Heinle Publishers.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 508-600). Washington, DC: American Council of Education.

Austin, J. L. (1962). *How to do things with words.* Oxford: Clarendon Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (1980). The construct validation of oral proficiency tests. *TESL Studies, 3*, 1-20.

Bachman, L. F., & Palmer, A. (1981a). A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In A. Palmer, P. Groot, & G. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 149-165). Washington, DC: TESOL.

Bachman, L. F., & Palmer, A. (1981b). The construct validation of the FSI oral interview. *Language Learning, 31*, 67-86.

Bachman, L. F., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449-465.

Bachman, L. F., & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*, 14-29.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bae, J., & Bachman, L. F. (1998). A latent variable approach to listening and reading: testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Language Testing, 15*, 380-414.

Barnwell, D. (1996). *A history of foreign language testing in the United States: From its beginnings to the present*. Tempe, AZ: Bilingual Press.

Becker, C. A. (1982). The development of semantic context effects: Two processes or two strategies? *Reading Research Quarterly, 17*, 482-502.

Beglar, D. (2000). *The validation of a breadth of vocabulary test using structural equation modeling*. Unpublished doctoral dissertation, Temple University, Philadelphia.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (2006). *EQS 6 Structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & D. G Bonett. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Bentler, P. M., & Dijkstra, T. (1985). Efficient estimation via linearization in structural models. In P. R. Krishnaiah (Ed.), *Multivariate analysis VI* (pp. 9-42). Amsterdam: North-Holland.

Bentler, P. M., & Weeks, D. G (1980). Linear structural equations with latent variables. *Psychometrika, 45*, 289-308.

Bernhardt, E. B. (1991). *Reading development in a second language: Theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex.

Bernhardt, E. B. (1999). If reading is reader-based, can there be a computer-adaptive test of reading? In M. Chalhoub-Deville, M (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 1-10). Cambridge, UK: University of Cambridge Local Examinations Syndicate.

Bernhardt, E. B., & Kamil, M. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics, 16*, 15-34.

Birch, B. (2007). *English L2 reading: Getting to the bottom*. Mahwah, NJ: Erlbaum Associates.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Kratwohl, D. R. (1956). Taxonomy of educational objectives: cognitivedomain. New York: David McKay.

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Bossers, B. (1991). On thresholds, ceilings and short-circuits: The relationship between L1 reading, L2 reading and L2 knowledge. In J. H. Hulstijin & J. F. Mater (Eds.), *AILA Review, 8*, 45-60.

Brière, E. (1971). Are we really measuring proficiency with our foreign language tests? *Foreign Language Annuals, 4*, 385-391.

Briggs, P., Austin, S., & Underwood, G. (1984). The effects of sentence context in good and poor readers: A test of Stanovich's interactive-compensatory model. *Reading Research Quarterly, 20*, 54-61.

Brisbois, J. (1995). Connections between first-and second-language reading. *Journal of Reading Behaviour, 27*, 565-584.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Byrne, B. M. (2006). *Structural equation modeling with EQS* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Canale, M. (1983a). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Canale, M. (1983b). On some dimensions of language proficiency. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 333-342). Rowley, MA: Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Carrell, P. L. (1991). Second language reading: Reading ability or language proficiency? *Applied Linguistics, 12*, 159-179.

Carroll, J. (1961). *Fundamental considerations in testing for English language proficiency of foreign students.* Washington, DC: Center for Applied Linguistics of the Modern Language Association of America.

Carroll, J. (1968). Psychology of language testing. In A. Davies (Ed.), *Language testing symposium: a psycholinguistic approach* (pp. 46-69). London: Oxford University Press.

Carroll, J. (1983). Psychometric theory and language testing. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, MA: Newbury House.

Carroll, J. B. (1993). Human cognitive abilities. Cambridge: Cambridge University Press.

Celce-Murcia, M., & Larsen-Freemen, D. (1999). *The grammar book: An ESL/EFL teacher's course* (2nd ed.). Boston, MA: Heinle & Heinle.

Chang, J. (2004). *Examining models of second language knowledge with specific reference to relative clauses: a model-comparison approach.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clarke, M., & Silberstein, S. (1977). Toward a realization of psycholinguistic principles for the ESL reading class. *Language Learning, 27*, 135-154.

Crystal, D. (1997). *The Cambridge encyclopedia of language*. Cambridge: Cambridge University Press.

Daly, J. A., & Miller, M. D. (1975a). The empirical development of an instrument to measure writing apprehension. *Research in the Teaching of English, 9*, 242-249.

Daly, J. A., & Miller, M. D. (1975b). Further studies on writing apprehension: SAT scores, success expectations, willingness to take advanced courses and sex difference *Research in the Teaching of English, 9*, 250-256.

Davidson, F., & Henning, G. (1985). A self-rating scale of English proficiency: Rasch scalar analysis of items and rating categories. *Language Testing, 2*, 164-179.

Day, R., & Bamford, J. (1998). *Extensive reading in the second language classroom.* Cambridge: Cambridge University Press.

DeCarlo, L. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292-307.

Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first and second language learners. *Reading Research Quarterly, 38*, 78-103.

English Language Institute (2006a). *Michigan Certificate Examinations General Information Bulletin 2005-2006.* Ann Arbor, MI: English Language Institute, The University of Michigan.

English Language Insitute (2006b). *The ECPE Annual Report: 2003-04.* Ann Arbor, MI: English Language Institute, The University of Michigan.

English Language Insitute (2006c). *The ECPE Annual Report: 2004-05.* Ann Arbor, MI: English Language Institute, The University of Michigan.

Farhady, H. (1983). New directions for ESL proficiency testing. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 253-269). Rowley, MA: Newbury House.

Felan, G. D. (2002, February). Test equating: mean, liner, equipercentile, and item response theory. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fouly, K. (1985). *A confirmatory multivariate study of the nature of second language proficiency and its relationships to learner variables.* Unpublished doctoral dissertation, University of Illinois, Urbana.

Gao, H. (2004). *The effect of different anchor tests on the accuracy of test equating for test adaptation.* Unpublished doctoral dissertation, Ohio University, Athens.

Garson, D. (n.d.). Statnotes: Topics in multivariate analysis. Retrieved February 23, 2004, from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm

Ginther, A., & Stevens, J. (1998). Language background, ethnicity, and the internal construct validity of the advanced placement Spanish language examination. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 169-194). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Goodman, K. S. (1967). Reading: a psycholinguistic guessing game. *Journal of the Reading Specialist, 6*, 126-135.

Goodman, K. S. (1969). Analysis of oral reading miscues: Applied psycholinguistics. *Reading Research Quarterly, 5*, 9-30.

Goodman, K. S. (1976). Reading: A psycholinguistic guessing game. In H. Singer & R. B. Ruddell (Eds.), Theoretical models and processes of reading (2nd ed., pp. 497-508). Newark: DE: International Reading Association.

Gough, P. B. (1985). One second of reading. In H. Singer & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (3rd ed., pp. 661-686). Newark: DE: International Reading Association.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*, 375-406.

Grabe, W. (1999). Developments in reading research and their implications for computer-adaptive reading assessment. In M. Chalhoub-Deville, M (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 11-47). Cambridge, UK: University of Cambridge Local Examinations Syndicate.

Grabe, W. (2005). The role of grammar in reading development. In J. Frodesen & C. Holton (Eds.), *The power of context in language teaching and learning* (pp. 129-139). Boston: Heinle & Heinle.

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading: Applied linguistics in action*. New York: Longman.

Gujarati, D. N. (1988). *Basic econometrics* (2nd ed.). New York: McGraw-Hill, Inc.

Halliday, M. (1973). *Explorations in the functions of language.* London: Edward Arnold.

Halliday, M. (1976). The form of a functional grammar. In G. Kress (Ed.), *Halliday: System and function in language.* Oxford: Oxford University Press.

Halliday, M., McIntosh, A., & Strevens, P. (1964). *The linguistic sciences and language teaching.* Bloomington: Indiana University Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. London: SAGE Publications.

Harley, B., Cummins, J., Swain, M., & Allen, P. (1990). The nature of language proficiency. In B. Harley, J. Cummins, M. Swain, & P. Allen (Eds.), *The development of second language proficiency* (pp. 7-25). Cambridge: Cambridge University Press.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle & Heinle Publishers.

Hedge, T. (2000). *Teaching and learning in the language classroom*. Oxford: Oxford University Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth, UK: Penguin.

Irvine, P., Atai, P., & Oller, J. W. (1974). Cloze, dictation, and the test of English as a foreign language. *Language Learning, 24*, 245-252.

Johnson, J., Yamashiro, A., & Yu, J. (2003). *ECPE annual report: 2002*. Ann Arbor, MI: English Language Institute, University of Michigan.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Lang (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it.* Newbury Park, CA: Sage University Press.

Kim, J. S., & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26*, 255-270.

Kline, R. B. (1998). *Principles and practices of structural equation modeling*. New York: The Guilford Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: methods and practices* (2nd ed.). New York: Springer.

Kunnan, A. J. (1995). *Test-taker characteristics and test performance: A structural equation modeling approach.* Cambridge: Cambridge University Press.

Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing, 15*, 295-332.

LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293-323.

Lado, R. (1961). *Language testing*. New York: McGraw Hill.

Larsen-Freeman, D. (1991). Teaching grammar. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 279 – 296). Boston: Heinle & Heinle.

Lee, S. (2005). Facilitating and inhibiting factors in English as a foreign language writing performance: A model testing with structural equation modeling. *Language Learning, 55*, 335-374.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*, 357-372.

Leech, G. (1983). *Principles of pragmatics*. London: Longman.

Lennon, R. T. (1962). What can be measured? *Reading Teacher, 15*, 326-337.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Lunzer, E., & Gardner, K. (Eds.). (1979). *The effective use of reading*. London: Heinemann Educational Books.

Luo, G., Seow, A., & Chin, C. L. (2001). Linking and anchoring techniques in test equating using the Rasch model. Retrieved Feb 23, 2004, from http://dspace.lboro.ac.uk/dspace/handle/2134/1817

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*, 519-530.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya, B36*, 115-128.

McDonald, R. P., & Marsh, H. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 107*, 247-255.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd. ed., pp. 13-103). New York: Oryx Press.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Murtagh, L. (1989). Reading in a second or foreign language: models, processes, and pedagogy. *Language, Culture and Curriculum, 2*, 91-105.

Oller, J. W. (1979). *Language tests at school*. London: Longman Group Limited.

Oller, J. W. (1983a). Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 3-10). Rowley, MA: Newbury House.

Oller, J. W. (1983b). A consensus for the eighties? In J. W. Oller (Ed.), *Issues in language testing research* (pp. 351-356). Rowley, MA: Newbury House.

Oller, J. W., & Hinofotis, F. B. (1980). Two mutually exclusive hypotheses about second language ability: Factor analytic studies of a variety of language subtests. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 13-23). Rowley, MA: Newbury House.

Park, T. J. (2007). *Investigating the construct validity of the community language program (CLP) English writing test.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Perfetti, C. A., Goldman, S. R., & Hogaboam, T. W. (1979). Reading skill and the identification of words in discourse context. *Memory and Cognition, 7*, 273-282.

Perfetti, C. A., & Lesgold, A. M. (1977). Discourse comprehension and sources of individual differences. In M. Just & P. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 141-183). Hillsdale, NJ: Erlbaum.

Perfetti, C. A., & Lesgold, A. M. (1979). Coding and comprehension in skilled reading and implications for reading instruction. In L. B. Resnick & P. A. Weaver (Eds.), *Theory and practice of early reading* (Vol. 1, pp. 57-85). Hillsdale, NJ: Erlbaum.

Porter, D. (1983). The effects of quantity of context on the ability to make linguistic predictions: a flaw in a measure of general proficiency. In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 63-74). London: Academic Press.

Purcell, E. T. (1983). Models of pronunciation accuracy. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 133-153). Rowley, MA: Newbury House.

Purpura, J. E. (1997). An analysis of the relationships between test taker's cognitive and metacognitive strategy use and second language performance. *Language Learning, 47*, 289-325.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: A structural equation modeling approach. *Language Testing, 15*, 333-379.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach. Studies in Language Testing 8.* Cambridge: Cambridge University Press.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Pyo, K. H. (2001). Construct validation of an integrated-approach EAP placement test using multi-group structural equation modeling. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Radach, R., Kennedy, A., & Rayner, K. (2004). *Eye movements and information processing during reading*. Hove, UK: Psychology Press.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice Hall.

Rea-Dickins, P. (1991). What makes a grammar test communicative? In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 112-131). London: MacMillan.

Rose, M. (1984). *Writer's block: The cognitive dimension.* Carbondale: Southern Illinois University Press.

Rumelhart, D. E. (1985). Toward an interactive model of reading. In H. Singer & R. B. Ruddell (Eds.), Theoretical models and processes of reading (3rd ed., pp. 722-750). Newark: DE: International Reading Association.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the examination for the certificate of proficiency in English. *Spaan Fellow Working papers in Second or Foreign Language Assessment, 1*, 39-82.

Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude and intelligence: A structural equation modeling approach. *Language Learning, 43*, 313-344.

Sasaki, M. (1999). *Second language proficiency, foreign language aptitude, and intelligence*. Cambridge: Cambridge University Press.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Scholz, G., Hendricks, D., Spurling, R., Johnson, M., & Vandenberg, L. (1980). Is language ability divisible or unitary? A factor analysis of 22 English language proficiency tests. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 24-33). Rowley, MA: Newbury House.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing, 22*, 1-30.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language.* Cambridge: Cambridge University Press.

Segalowitz, N., Poulsen, C., & Komoda, M. (1991). Lower level components of reading skill in higher level bilinguals: Implications for reading instruction. *AILA Review, 8*, 15-30.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*, 99-128.

Skehan, P. (1991). Progress in language testing: The 1990s. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 3-21). London: MacMillan.

Smith, F. (1971). *Understanding reading*. New York: Holt, Rinehart & Winston.

Smith, F. (1982). *Understanding reading* (3th ed.). New York: Holt, Rinehart & Winston.

Smith, F. (1988). *Understanding reading* (4th ed.). New Jersey: Lawrence Erlbaum Associates.

Spolsky, B. (1968). Linguistics and language pedagogy – applications or implications? In J. E. Alatis (Ed.), *Twentieth annual round table on languages and linguistics* (pp. 143-155). Washington, DC: Georgetown University Press.

Spolsky, B. (1973). What does it mean to know a language; or how do you get someone to perform his competence? In J. W. Oller & J. Richards (Eds.), *Focus on the learner* (pp. 164-176). Rowley, Mass: Newbury House.

SPSS, Inc. (2003). SPSS statistics for research and analysis (Version 12.0) [Computer software]. Chicago, IL: SPSS, Inc.

Stanovich, K. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32-71.

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.

Stanovich, K. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford Press.

Stricker, L. J., Rock, D. A., & Lee, Y. W. (2005). *Factor structure of the LanguEdge™ test across language groups* (Monograph Series MS-32). New Jersey: Educational Testing Service.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Tsai, C. H. L. (2004). *Investigating the relationships between ESL writers' strategy use and their second language writing ability.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York.

Turner, C. E. (1989). The underlying factor structure of L2 cloze test performance in Francophone, university-level students: Casual modeling as an approach to construct validation. *Language Testing, 6*, 172-197.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. London: Addison Wesley Longman Limited.

Vollmer, H., & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 29-79). Rowley, MA: Newbury House.

Wagner, E. (2004). A construct validation study of the listening sections of the ECPE and MELAB. *Spaan Fellow Working papers in Second or Foreign Language Assessment, 2*, 1-25.

Whitney, P., & Budd., D. (1996). Think-aloud protocols and the study of comprehension. *Discourse Processes, 21*, 341-51.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention* (pp. 281-324). Washington DC: American Psychological Association.

Widdowson, H. G. (1978). *Teaching language communication.* Oxford: Oxford University Press.

Woldbeck, T. (1998, April). *Basic concepts in modern methods of test equating.* Paper presented at the annual meeting of the Southwest Psychological Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 417215)

Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing, 22*, 463-508.

Yuan, K. H., & Bentler, P. M. (2004). On chi square difference and z-tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*, 737-757.

Yun, Y. (2005). *Factors explaining EFL learners' performance in a timed essay writing test: a structural equation modeling approach.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.